Sustainability in AI

Roy Schwartz

Hebrew University of Jerusalem

Guest lecture, Machine Learning CS-433
December 2022

THE HEBREW
UNIVERSITY
OF JERUSALEM

A Little about Me

Roy Schwartz

- A senior lecturer at the School of CS at the Hebrew U. of Jerusalem
  - I was a postdoc and a research scientist at The University of Washington and the Allen Institute for AI (AI2)

A Little about Me

Roy Schwartz

- A senior lecturer at the School of CS at the Hebrew U. of Jerusalem
  - I was a postdoc and a research scientist at The University of Washington and the Allen Institute for AI (AI2)
- I study Artificial Intelligence (AI) and focus on Natural Language Processing (NLP)
  - **Understanding** AI models
  - Revealing **biases** in datasets
  - Making AI more **sustainable**

A Little about Me

Roy Schwartz

- A senior lecturer at the School of CS at the Hebrew U. of Jerusalem
  - I was a postdoc and a research scientist at The University of Washington and the Allen Institute for AI (AI2)
- I study Artificial Intelligence (AI) and focus on Natural Language Processing (NLP)
  - **Understanding** AI models
  - Revealing **biases** in datasets
  - Making AI more **sustainable**

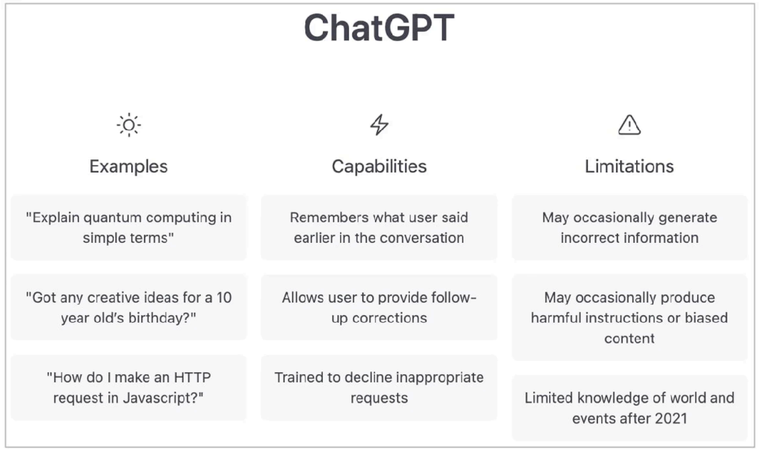
A Little about Me

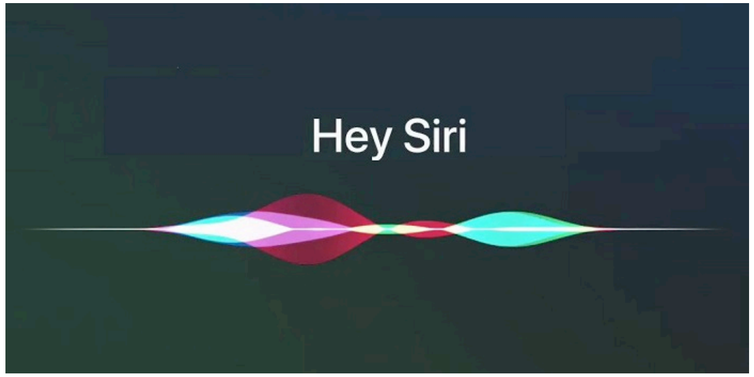
Roy Schwartz

- A senior lecturer at the School of CS at the Hebrew U. of Jerusalem
  - I was a postdoc and a research scientist at The University of Washington and the Allen Institute for AI (AI2)
- I study Artificial Intelligence (AI) and focus on Natural Language Processing (NLP)
  - **Understanding** AI models
  - Revealing **biases** in datasets
  - Making AI more **sustainable**

AI Today

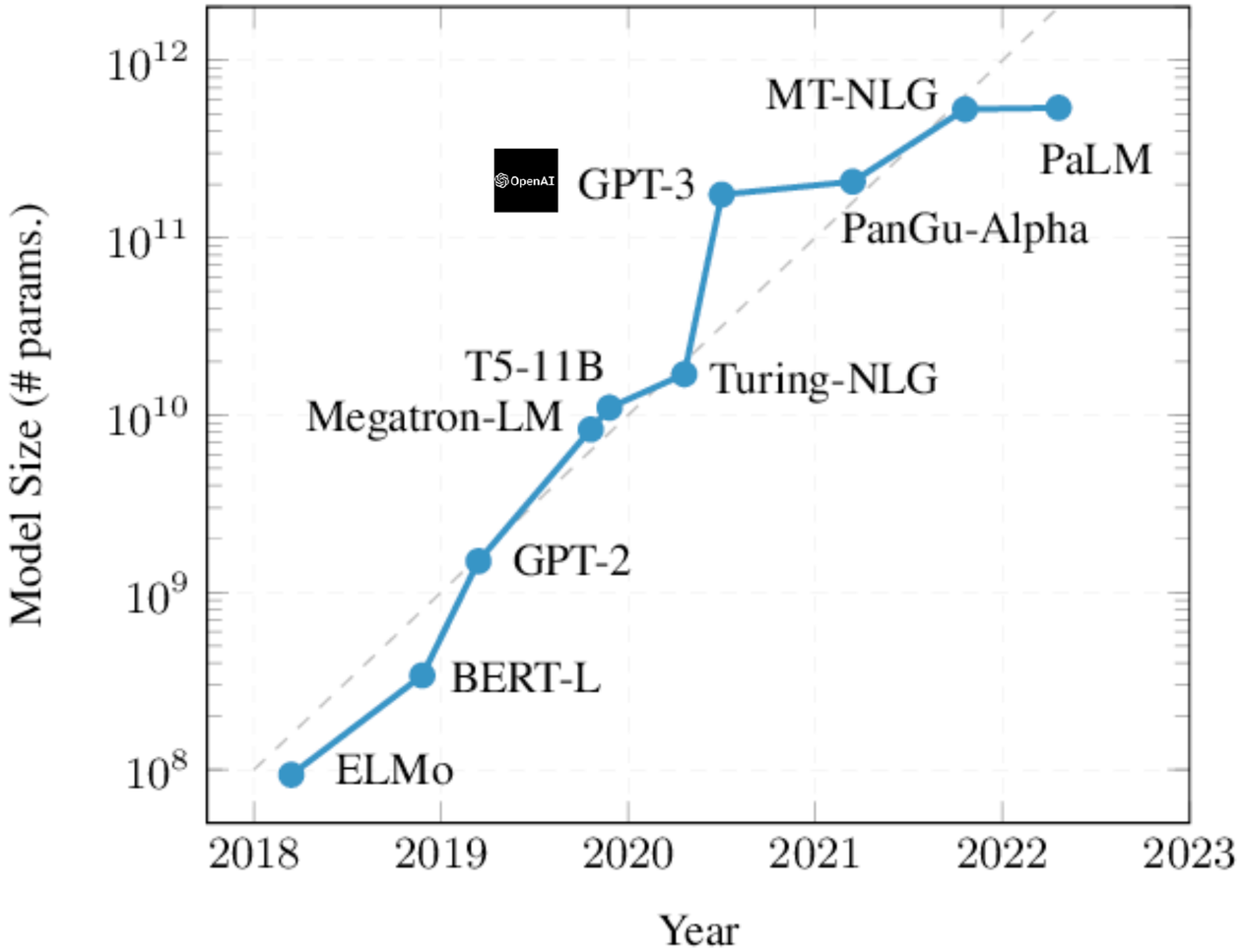
Microsoft Translator





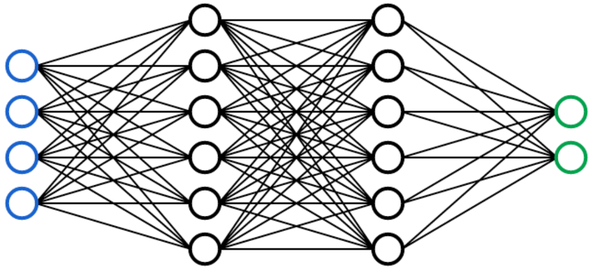
Large Models

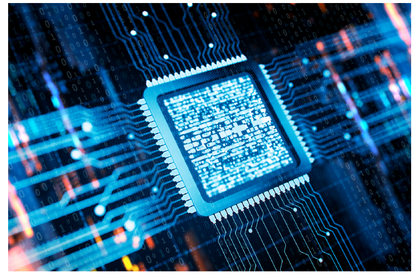
5,000X in 4 Years



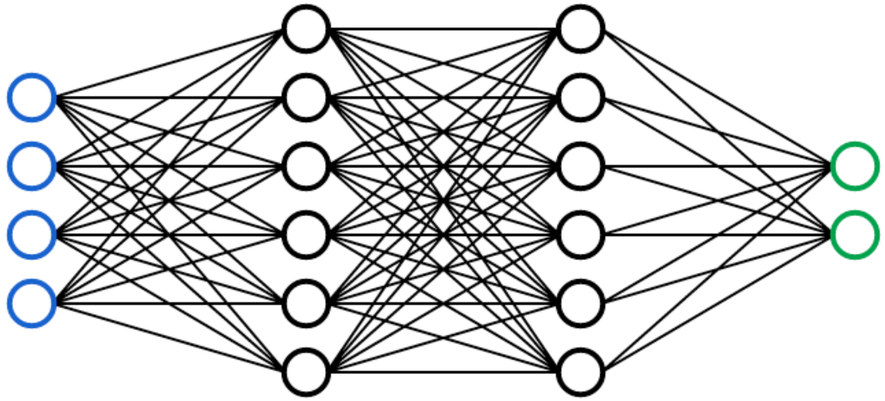
Taken from Lakim et al. (2022)

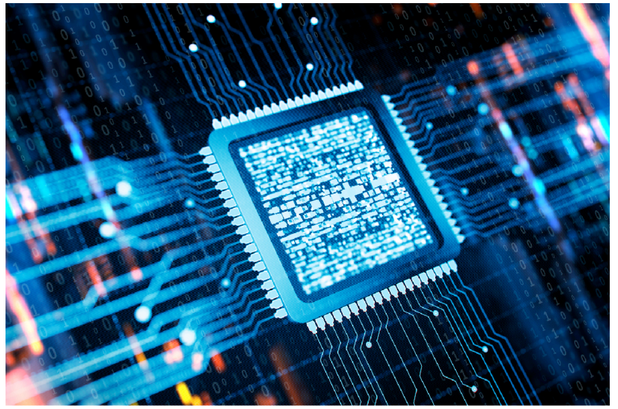
Scaling Neural Networks





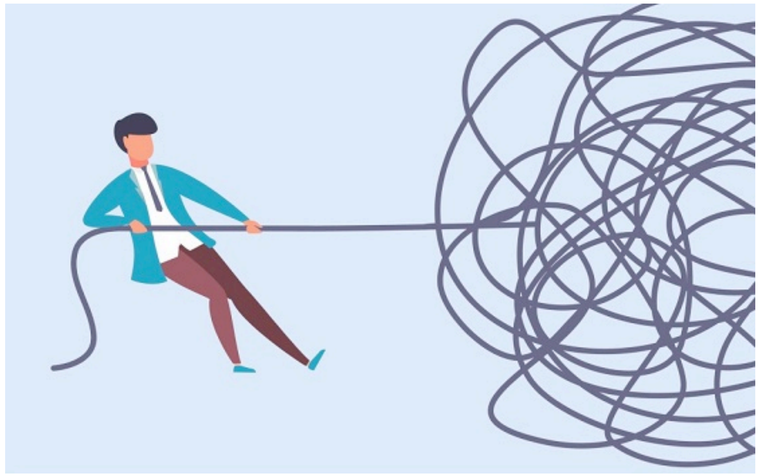
Scaling Neural Networks





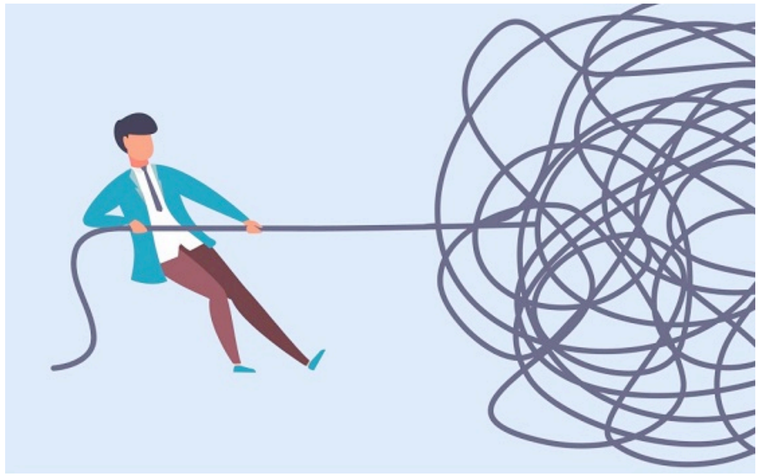
Outline

Outline

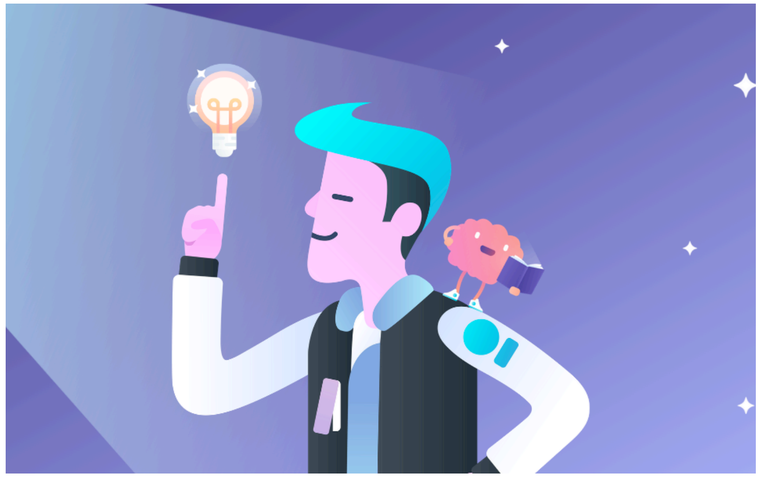


Challenges

Outline

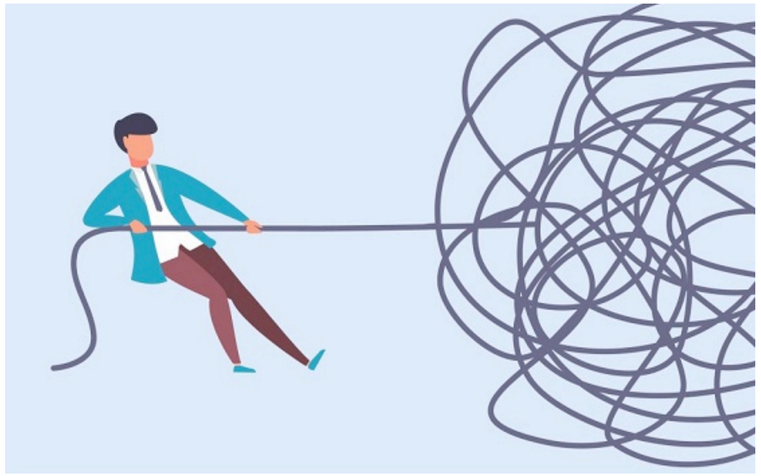


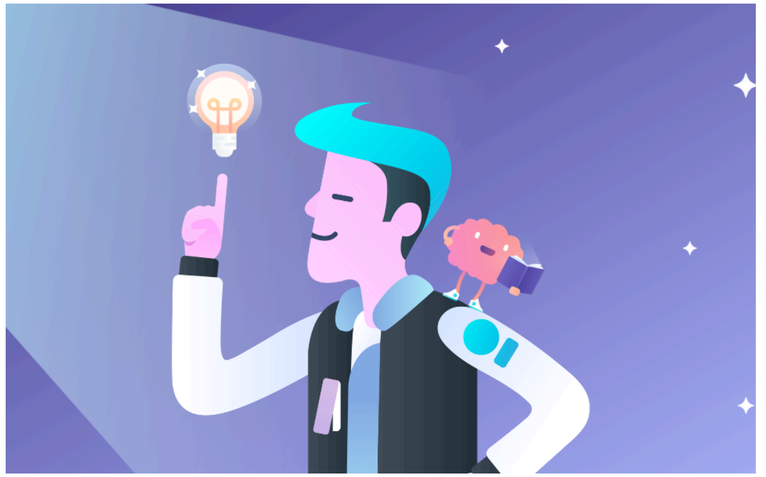
Challenges



Mitigation

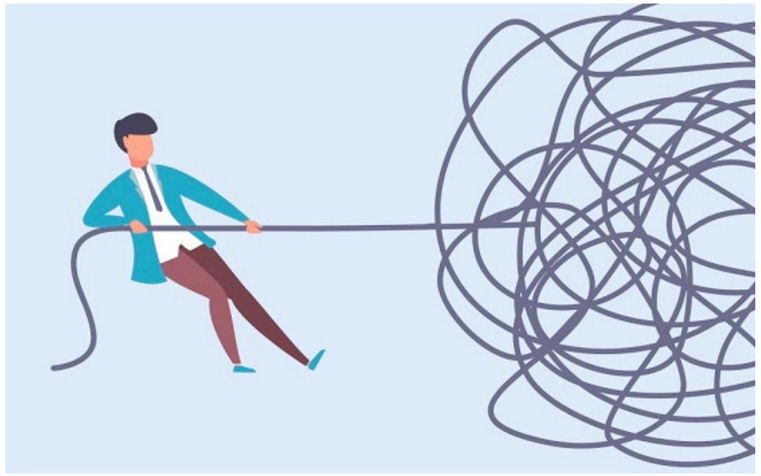
Outline



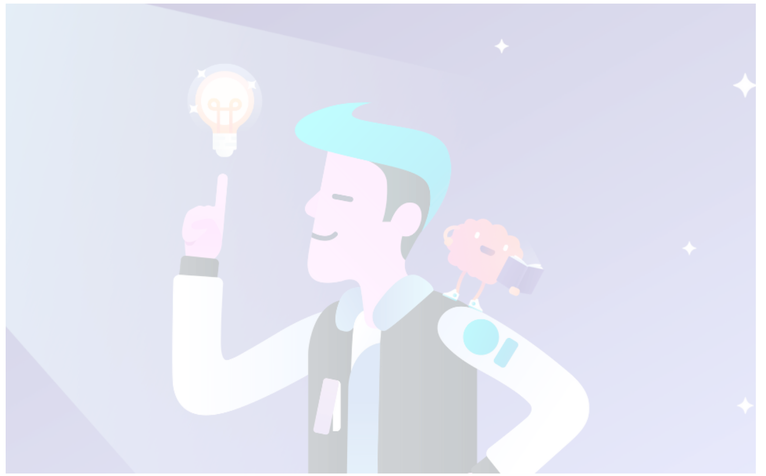


Challenges **Opportunities** **Mitigation**

Outline



Challenges



Mitigation

Opportunities

Challenges with Scaling

Inclusiveness

Synced

AI TECHNOLOGY & INDUSTRY REVIEW

FEATURE ▾ INDUSTRY ▾ TECHNOLOGY COMMUNITY ▾ ABOUT US ▾ REPORT CONTRIBUTE TO SYNCED REVIEW

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Training Costs

- BERT (Devlin et al, 2019) was trained on **16** Cloud TPUs for **4** days
- RoBERTa (Liu et al., 2019) was trained on **1024** V100 GPUs for approximately **1** day
- PaLM (Chowdhery et al., 2022) was trained on **6144** TPU v4 chips for **50** days and **3072** TPU v4 chips for **15** days

Number of Authors

Number of Authors

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

Number of Authors

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

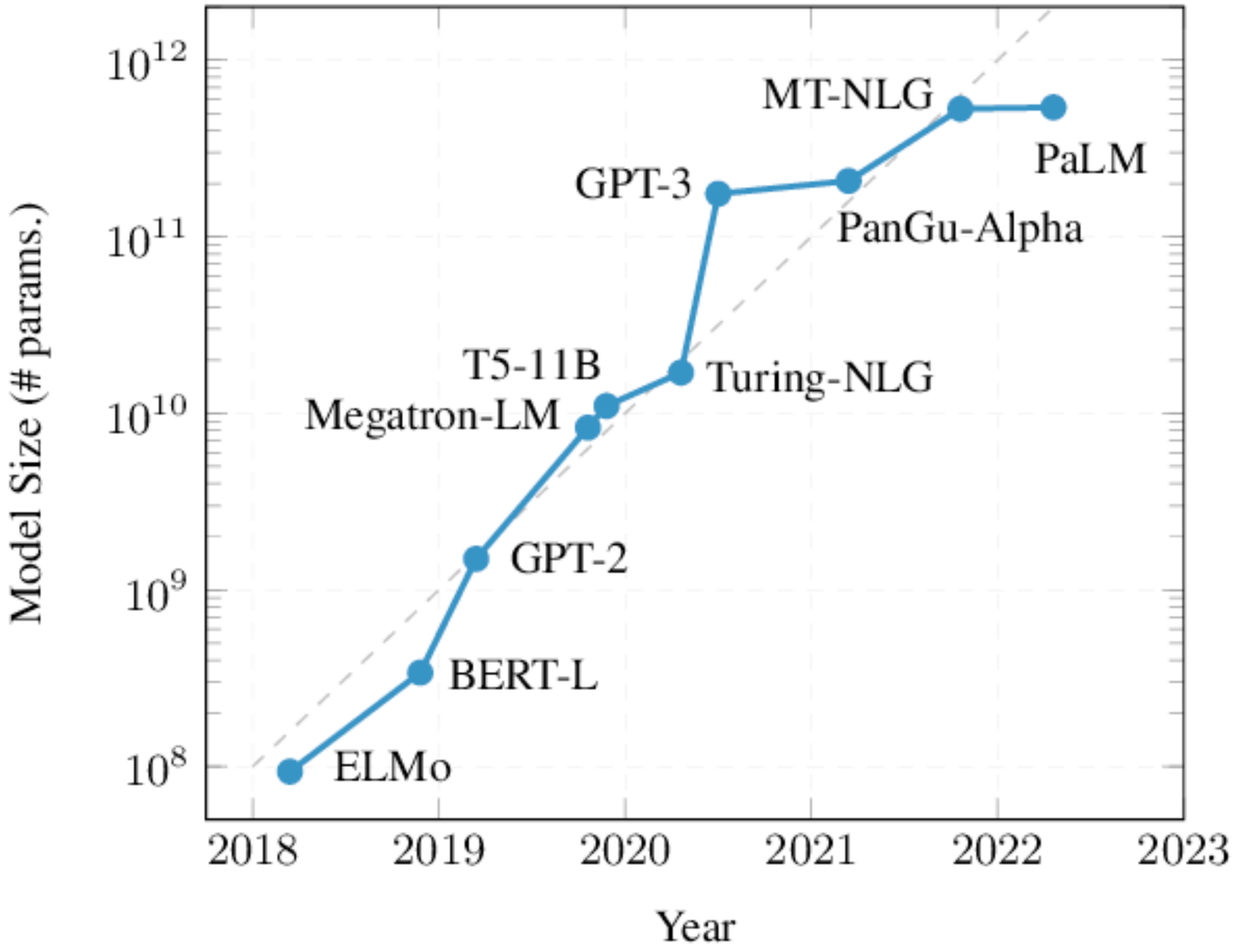
OpenAI

PaLM: Scaling Language Modeling with Pathways

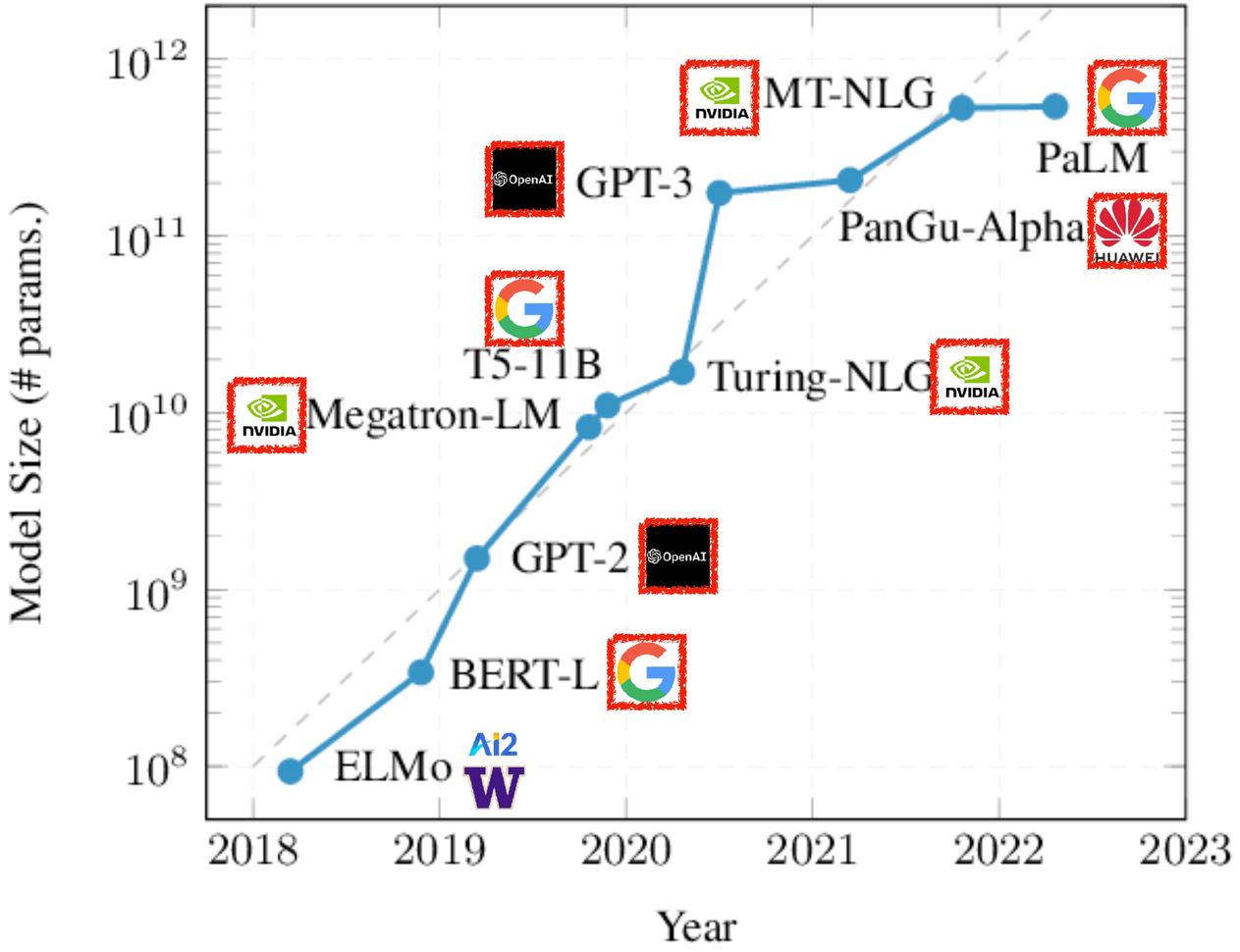
Aakanksha Chowdhery* Sharan Narang* Jacob Devlin*
Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham
Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi
Sasha Tsvyashchenko Joshua Maynez Abhishek Rao† Parker Barnes Yi Tay
Noam Shazeer‡ Vinodkumar Prabhakaran Emily Reif Nan Du Ben Hutchinson
Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari
Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev
Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus
Denny Zhou Daphne Ippolito David Luan‡ Hyeontaek Lim Barret Zoph
Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick
Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz
Erica Moreira Rewon Child Oleksandr Polozov† Katherine Lee Zongwei Zhou
Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta† Jason Wei
Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

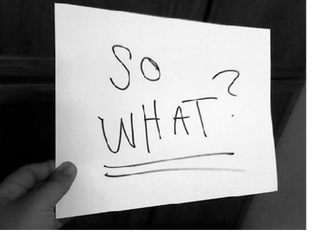
Google Research

It's a Rich Man's World



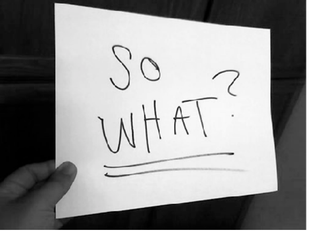
It's a Rich Man's World





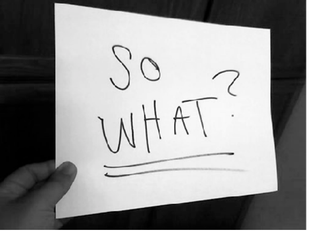
Lack of Inclusiveness

- Lack of diversity
  - Of opinions
  - Of cultures
  - Of backgrounds



Lack of Inclusiveness

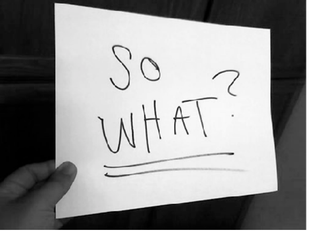
- Lack of diversity
  - Of opinions
  - Of cultures
  - Of backgrounds
- Who gets to decide
  - How does state-of-the-art look like?
  - What to work on?
  - Which data to use?



Lack of Inclusiveness

- Lack of diversity
  - Of opinions
  - Of cultures
  - Of backgrounds
- Who gets to decide
  - How does state-of-the-art look like?
  - What to work on?
  - Which data to use?

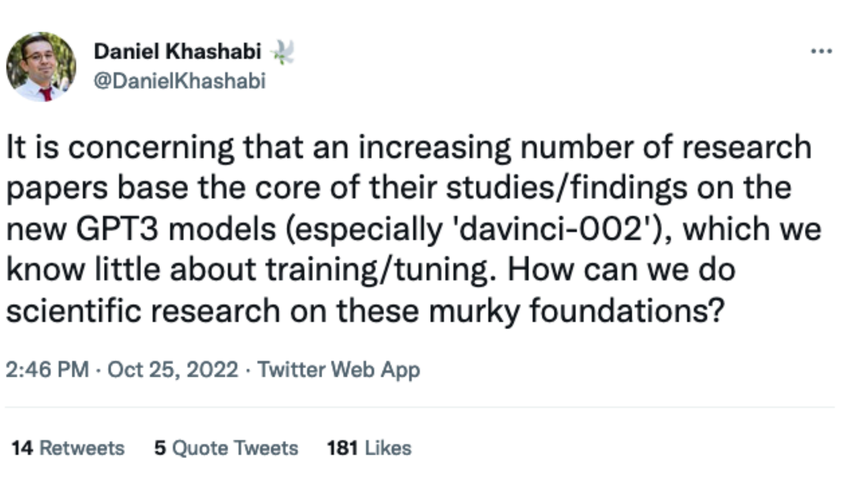
- Bad science!



Lack of Inclusiveness

- Lack of diversity
  - Of opinions
  - Of cultures
  - Of backgrounds
- Who gets to decide
  - How does state-of-the-art look like?
  - What to work on?
  - Which data to use?

- Bad science!



Energy and Policy Considerations for Deep Learning in NLP

ACL 2019

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Consumption	**CO_2e (lbs)**
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Energy and Policy Considerations for Deep Learning in NLP

ACL 2019

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Consumption	**CO_2e (lbs)**
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Is AI really creating an environmental problem?

Google’s Answer: No!

BLOG ›

Good News About the Carbon Footprint of Machine Learning Training

TUESDAY, FEBRUARY 15, 2022

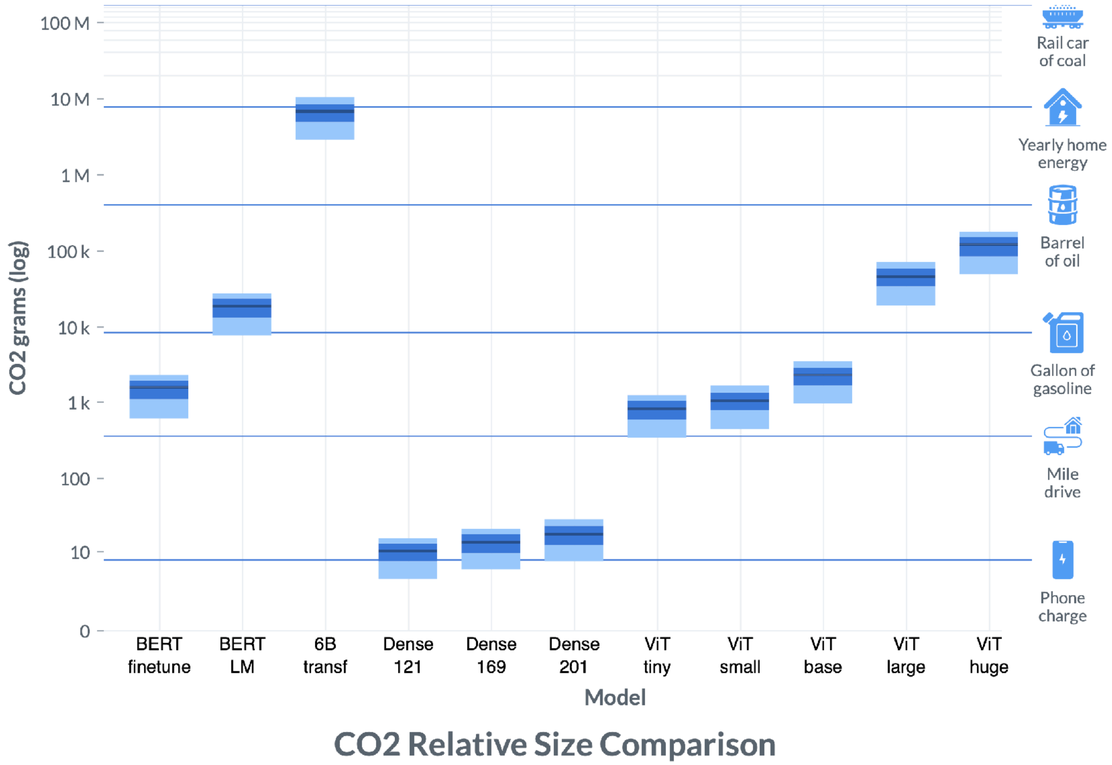
Posted by David Patterson, Distinguished Engineer, Google Research, Brain Team

*Strubell et al.’s energy estimate for NAS ended up **18.7X** too high for the average organization (…) and **88X** off in emissions for energy-efficient organizations like Google*

Our Answer: Maybe?

Measuring the Carbon Intensity of AI in Cloud Instances

JESSE DODGE, Allen Institute for AI, USA
TAYLOR PREWITT, University of Washington, USA
REMI TACHET DES COMBES, Microsoft Research Montreal, USA
ERIKA ODMARK, Microsoft, USA
ROY SCHWARTZ, Hebrew University of Jerusalem, Israel
EMMA STRUBELL, Carnegie Mellon University, USA
ALEXANDRA SASHA LUCCIONI, Hugging Face, USA
NOAH A. SMITH, Allen Institute for AI and University of Washington, USA
NICOLE DECARIO, Allen Institute for AI, USA
WILL BUCHANAN, Microsoft, USA



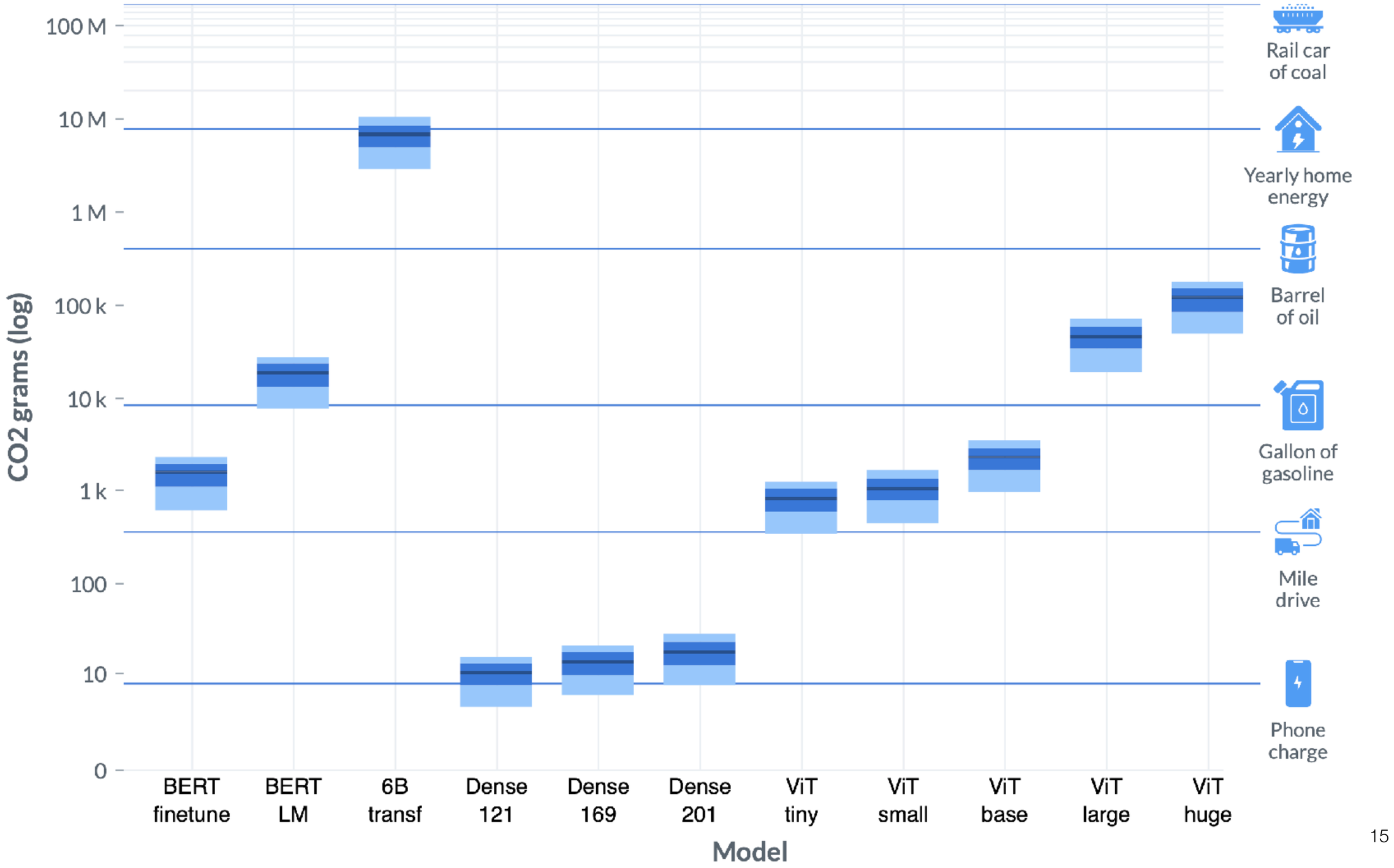
CO2 Relative Size Comparison

Green Software Foundation

Microsoft Azure

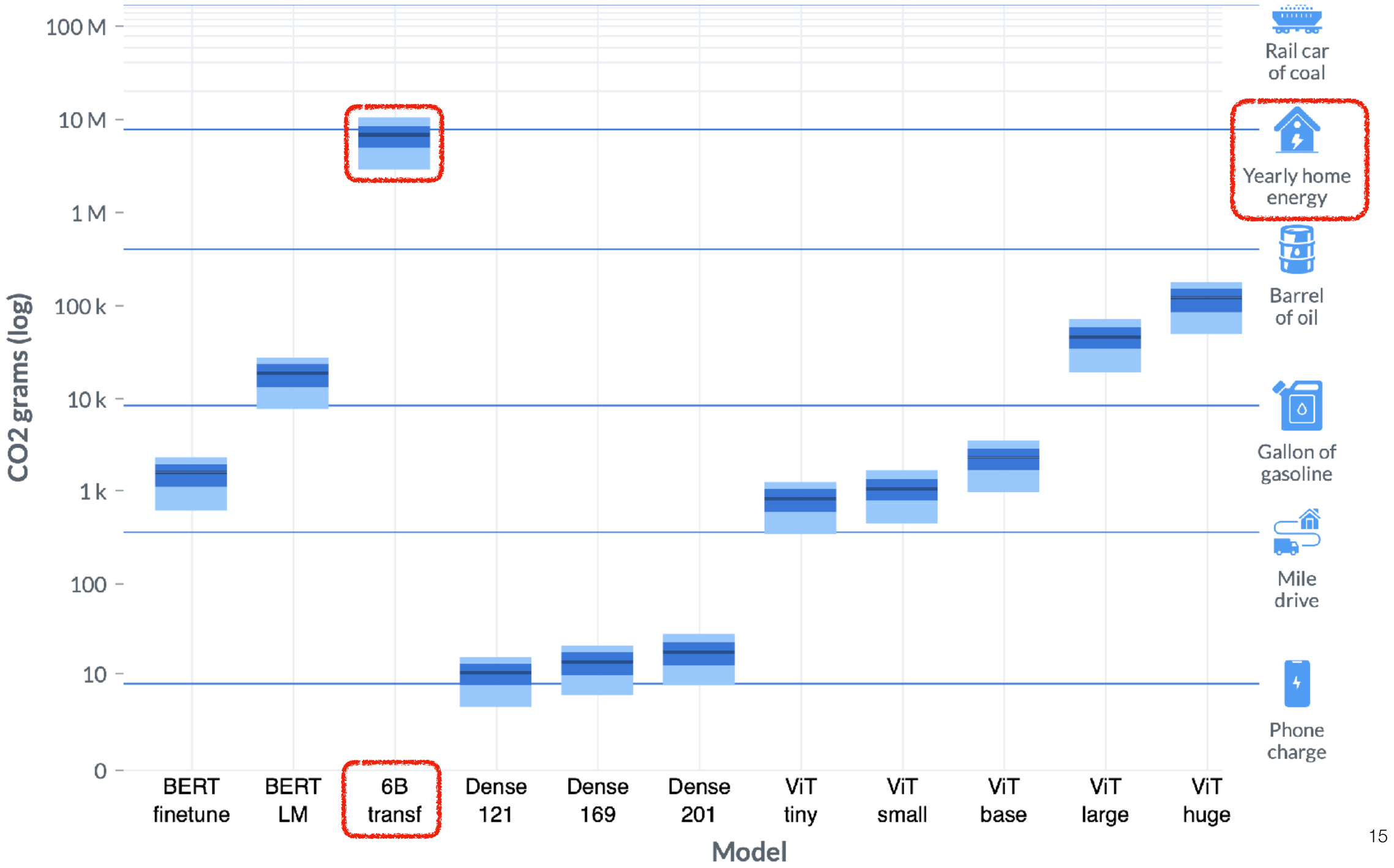
Measu

JESSE D
TAYLO
REMI T
ERIKA
ROY SC
EMMA
ALEXA
NOAH
NICOL
WILL B



Measu

JESSE D
TAYLO
REMI T
ERIKA
ROY SC
EMMA
ALEXA
NOAH
NICOL
WILL B



AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?

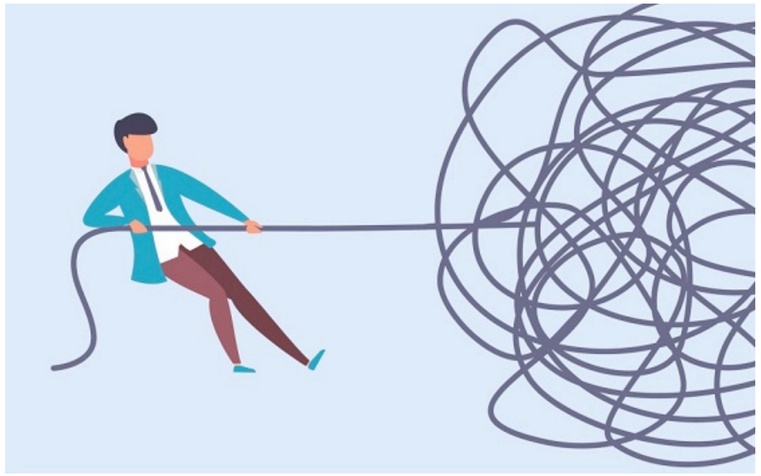
AI and the Environment

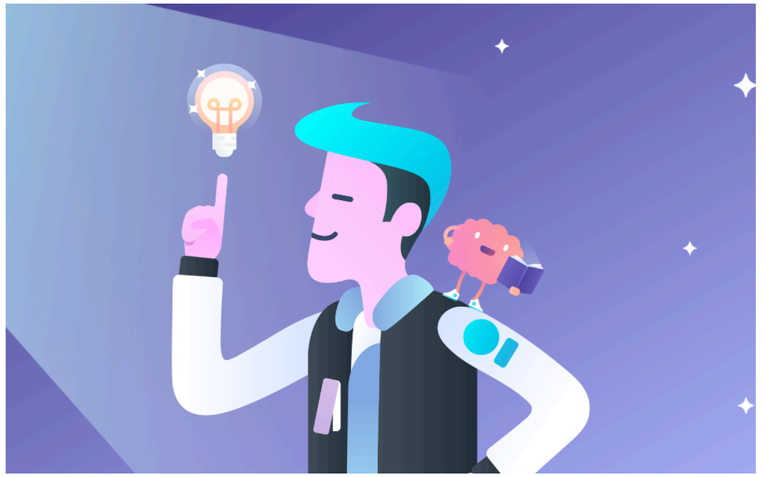
- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about inference operations?
  - Very cheap (though increasingly more expensive)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about inference operations?
  - Very cheap (though increasingly more expensive)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference
- Much harder to **measure**!

Outline



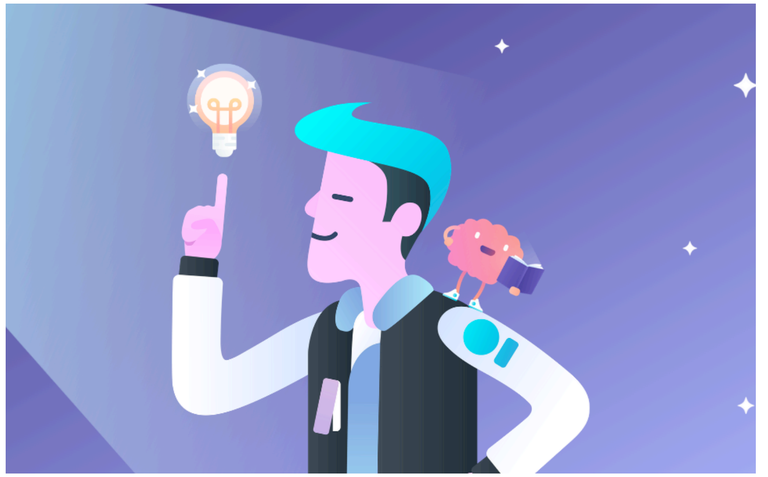


Challenges **Opportunities** **Mitigation**

Outline

Challenges

Opportunities



Mitigation

Stop training large models?

Large Models are Important

- Push the limits of SOTA
- Released large pre-trained models save compute
- Large models are potentially faster to train
  - Li et al. (2020)

Large Models are Important

- Push the limits of SOTA
- Released large pre-trained models save compute
- Large models are potentially faster to train
  - Li et al. (2020)
- But, large models have **concerning side affects**
  - Inclusiveness, environment

Large Models are Important

- Push the limits of SOTA
- Released large pre-trained models save compute
- Large models are potentially faster to train
  - Li et al. (2020)
- But, large models have **concerning side affects**
  - Inclusiveness, environment
- Our goal is to **mitigate these side affects**

Mitigating the Challenges of Scaling

Enhanced Reporting

Efficient Methods

Mitigating the Challenges of Scaling

Enhanced Reporting

Efficient Methods

Is Model A > Model B?

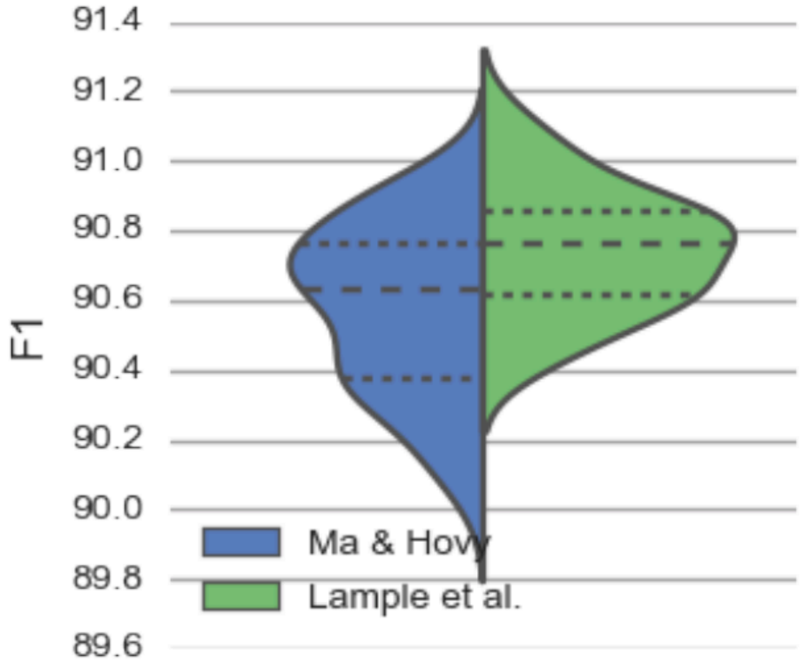
Reimers & Gurevych (2017)

Model	F1
Model A	**91.21**
Model B	90.94

Is Model A > Model B?

Reimers & Gurevych (2017)

Model	F1
Model A	**91.21**
Model B	90.94



Is Model A > Model B?

Melis et al. (2018)

Model	Size	Depth	Valid	Test
Medium LSTM, Zaremba et al. (2014)	10M	2	86.2	82.7
Large LSTM, Zaremba et al. (2014)	24M	2	82.2	78.4
VD LSTM, Press & Wolf (2016)	51M	2	75.8	73.2
VD LSTM, Inan et al. (2016)	9M	2	77.1	73.9
VD LSTM, Inan et al. (2016)	28M	2	72.5	69.0
VD RHN, Zilly et al. (2016)	24M	10	67.9	65.4
NAS, Zoph & Le (2016)	25M	-	-	64.0
NAS, Zoph & Le (2016)	54M	-	-	62.4
AWD-LSTM, Merity et al. (2017) †	24M	3	60.0	57.3

Perplexity (⬇)

Is Model A > Model B?

Melis et al. (2018)

Model	Size	Depth	Valid	Test
Medium LSTM, Zaremba et al. (2014)	10M	2	86.2	82.7
Large LSTM, Zaremba et al. (2014)	24M	2	82.2	78.4
VD LSTM, Press & Wolf (2016)	51M	2	75.8	73.2
VD LSTM, Inan et al. (2016)	9M	2	77.1	73.9
VD LSTM, Inan et al. (2016)	28M	2	72.5	69.0
VD RHN, Zilly et al. (2016)	24M	10	67.9	65.4
NAS, Zoph & Le (2016)	25M	-	-	64.0
NAS, Zoph & Le (2016)	54M	-	-	62.4
AWD-LSTM, Merity et al. (2017) †	24M	3	60.0	57.3

Perplexity (↓)

Is Model A > Model B?

Melis et al. (2018)

Model	Size	Depth	Valid	Test
Medium LSTM, Zaremba et al. (2014)	10M	2	86.2	82.7
Large LSTM, Zaremba et al. (2014)	24M	2	82.2	78.4
VD LSTM, Press & Wolf (2016)	51M	2	75.8	73.2
VD LSTM, Inan et al. (2016)	9M	2	77.1	73.9
VD LSTM, Inan et al. (2016)	28M	2	72.5	69.0
VD RHN, Zilly et al. (2016)	24M	10	67.9	65.4
NAS, Zoph & Le (2016)	25M	-	-	64.0
NAS, Zoph & Le (2016)	54M	-	-	62.4
AWD-LSTM, Merity et al. (2017) †	24M	3	60.0	57.3
LSTM	10M	1	61.8	59.6
LSTM		2	63.0	60.8
LSTM		4	62.4	60.1
RHN		5	66.0	63.5
NAS		1	65.6	62.7
LSTM	24M	1	61.4	59.5
LSTM		2	62.1	59.6
LSTM		4	60.9	58.3
RHN		5	64.8	62.2
NAS		1	62.1	59.7

Perplexity (↓)

Carefully Tuned (1500 trails)

Unfair Comparison

Is **Model A > Model B**?

Better(?) Comparison

Is **Model A > Model B?** | ***Budget***

Expected Validation

Dodge, Gururangan, Card, S. & Smith, EMNLP 2019

- Input: a set of experimental results $\{V_1, \ldots, V_n\}$
- Define $V_k^* = max_{i \in \{1,\ldots,k\}} V_i$

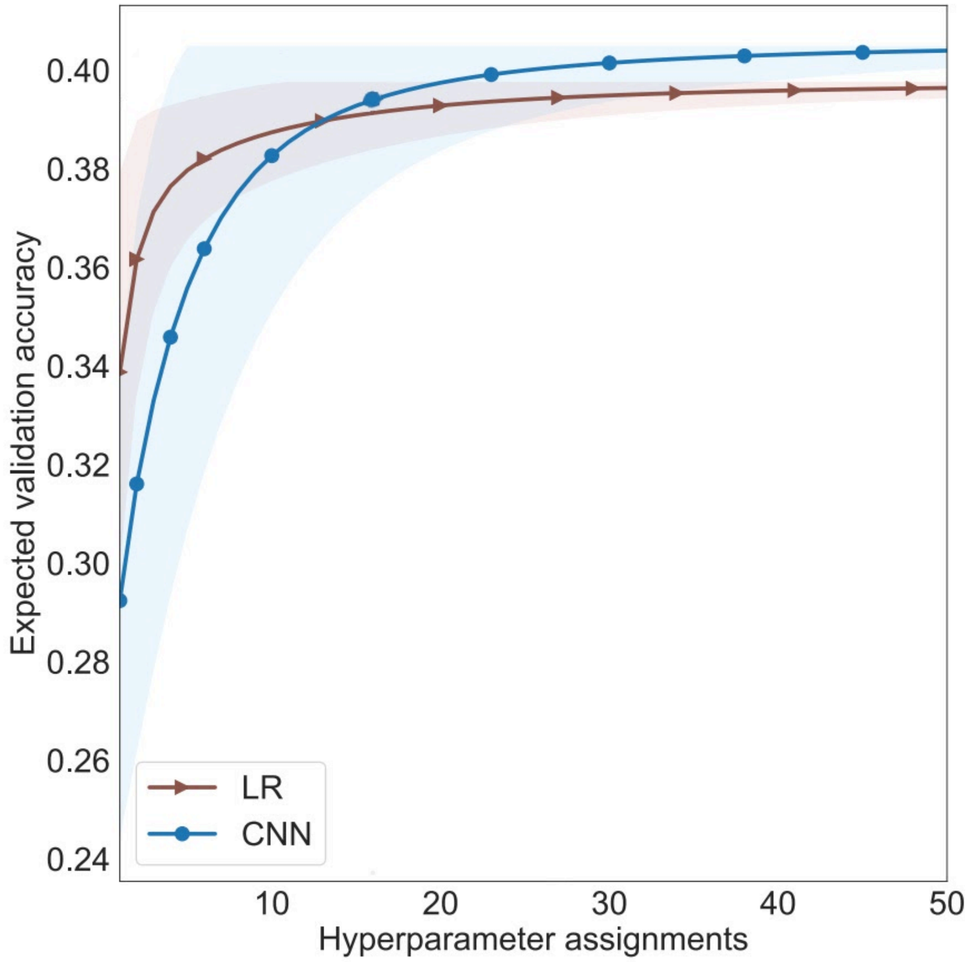
Expected Validation

Dodge, Gururangan, Card, S. & Smith, EMNLP 2019

- Input: a set of experimental results $\{V_1, \ldots, V_n\}$
- Define $V_k^* = max_{i \in \{1,\ldots,k\}} V_i$
- **Expected validation performance**: $\mathbb{E}[V_k^* \mid k]$
  - k=1: $mean(\{V_1, \ldots, V_n\})$
  - k=2: $mean(\{max(V_i, V_j) \forall 1 \leq i < j \leq n\})$
  - k=n: $V_n^* = max_{i \in \{1,\ldots,n\}} V_i$

Expected Validation

Example

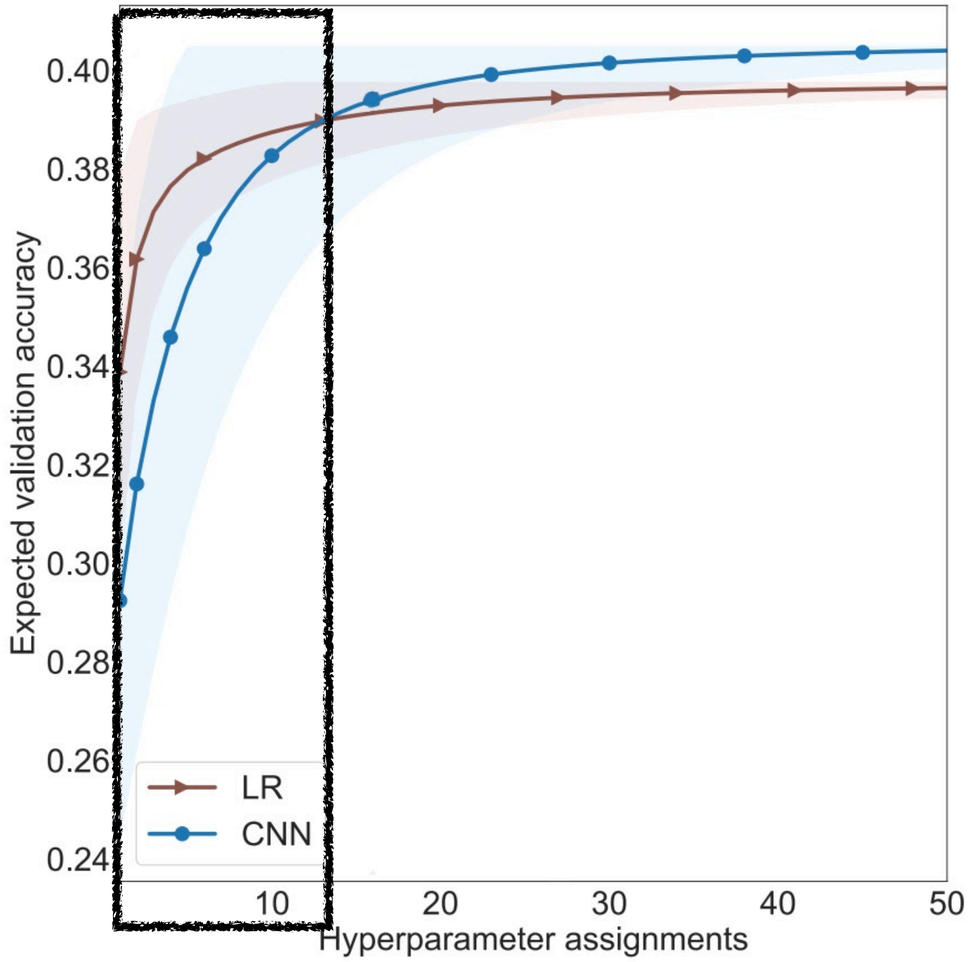


Expected Validation

Example

Low budget:

LR > CNN

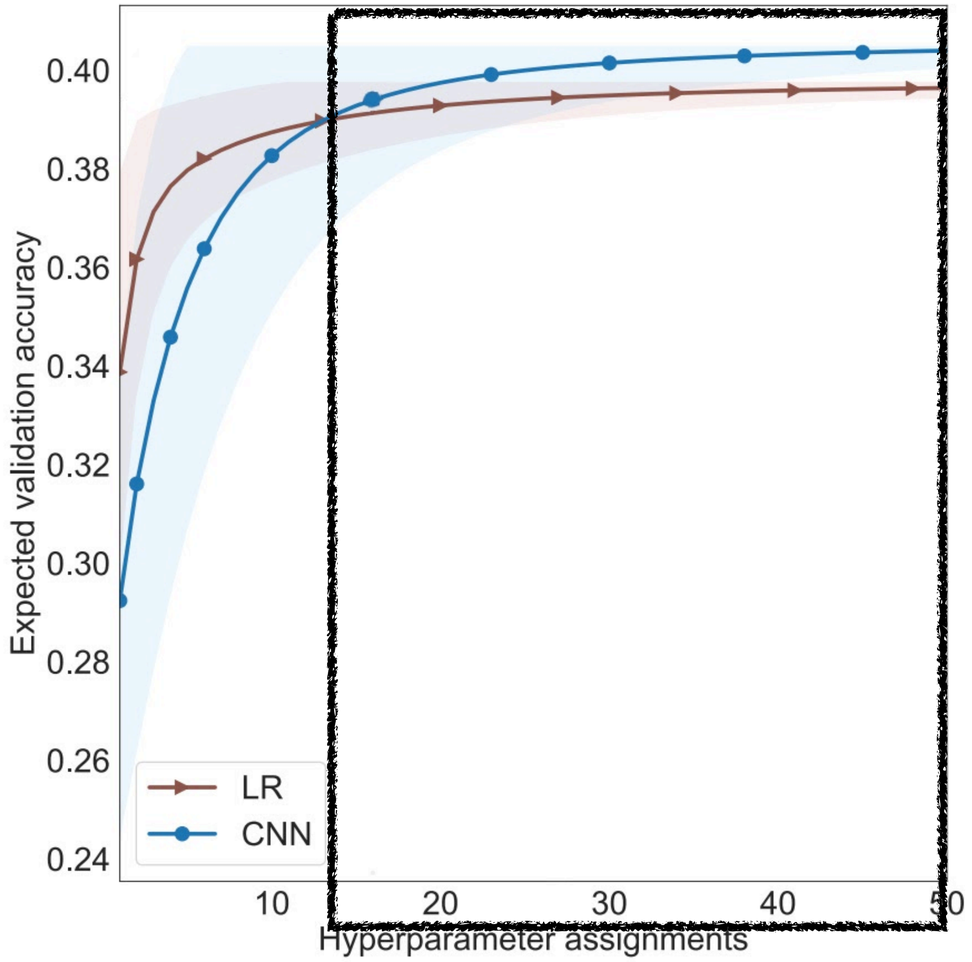


Expected Validation

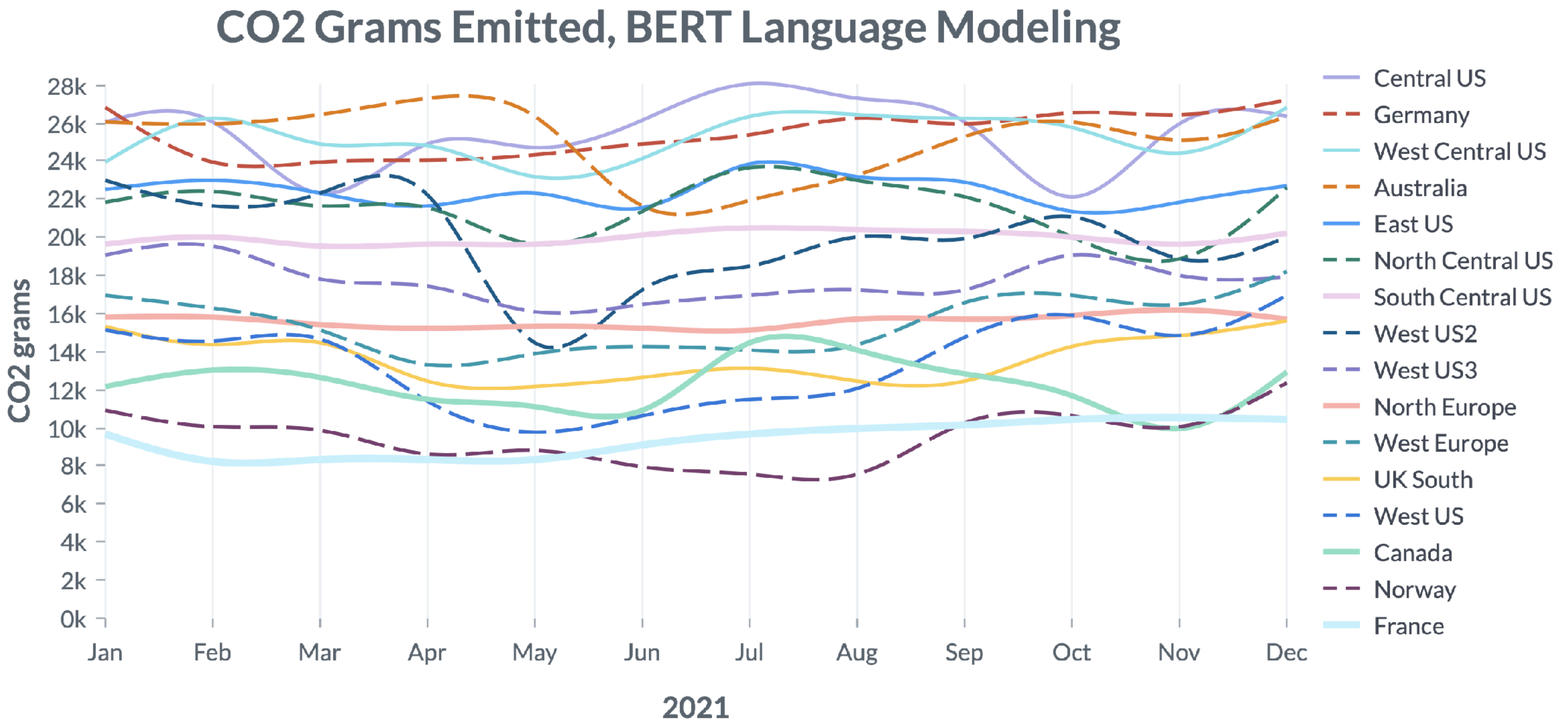
Example

High budget:

CNN > LR

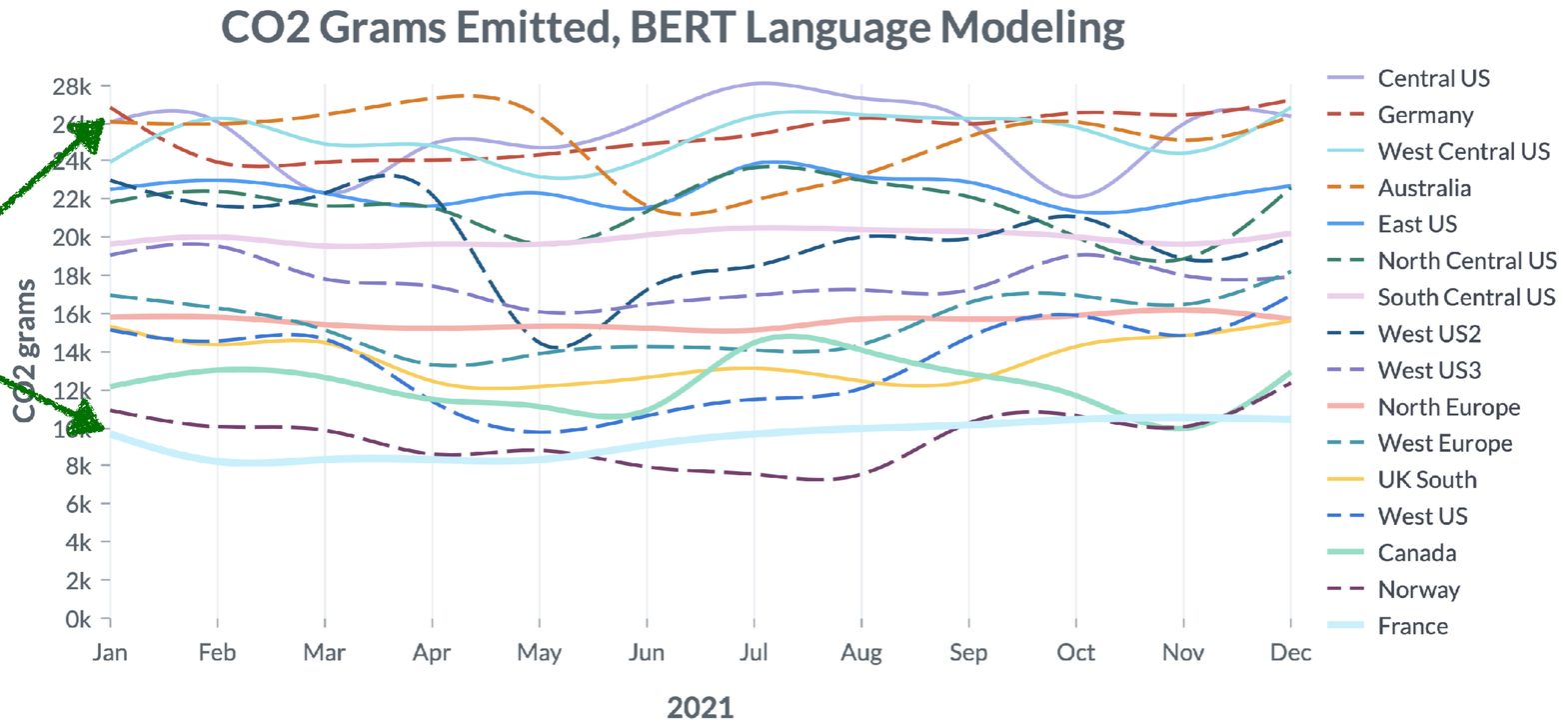


Cloud Location Matters



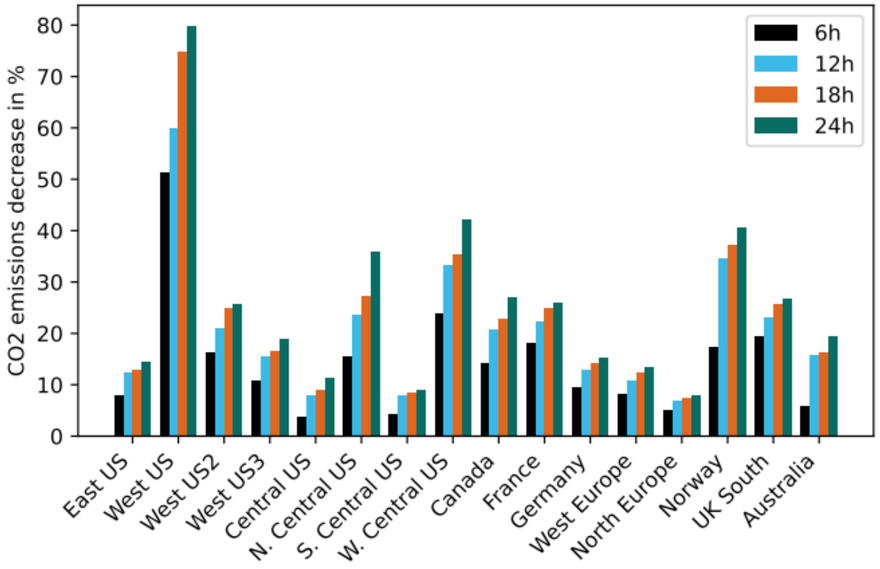
Cloud Location Matters

~3x difference

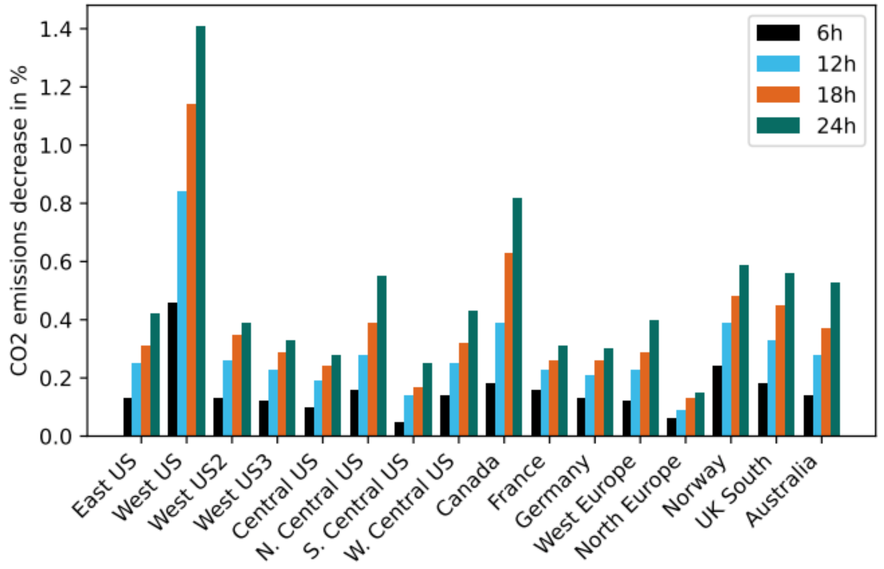


Time of Day Matters

Potential Saving with *Flexible Start*



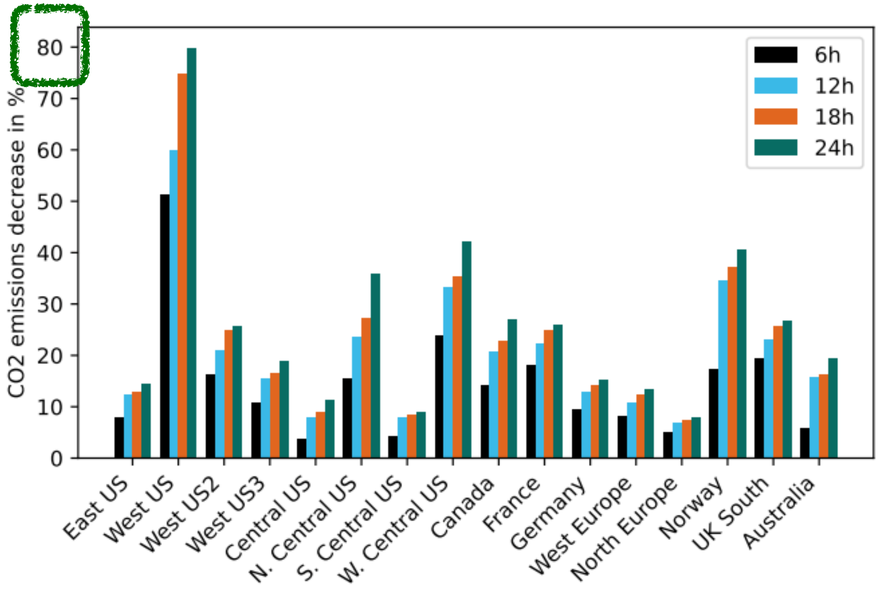
(a) *Flexible Start* optimization for Dense 201.



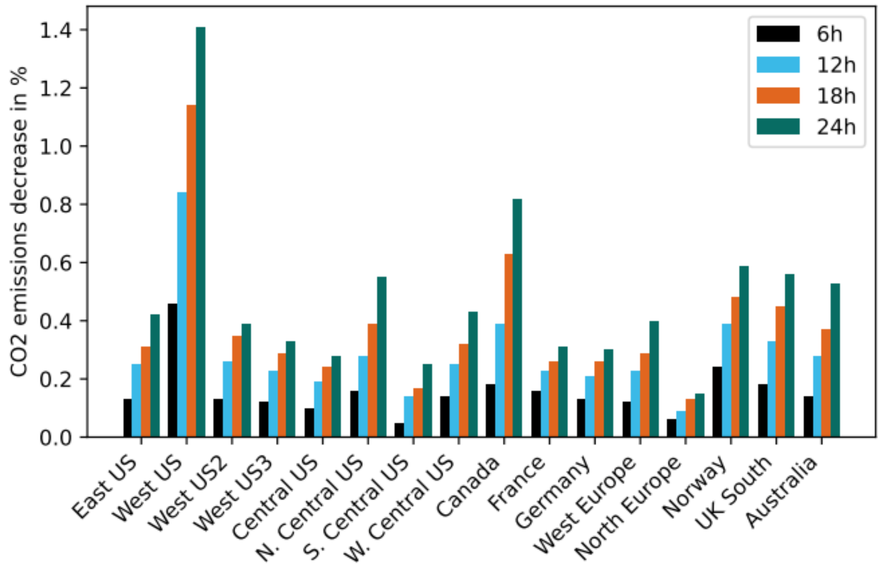
(b) *Flexible Start* optimization for 6B parameters Transformer.

Time of Day Matters

Potential Saving with *Flexible Start*



(a) *Flexible Start* optimization for Dense 201.



(b) *Flexible Start* optimization for 6B parameters Transformer.

Reporting
Open Questions

- How much will we gain by pouring **more compute**?

Reporting
Open Questions

- How much will we gain by pouring **more compute**?
- What should we report?
  - Number of experiments
  - Time
  - FLOPs
  - Energy (KW)
  - Carbon?

Mitigating the Challenges of Scaling

Enhanced Reporting

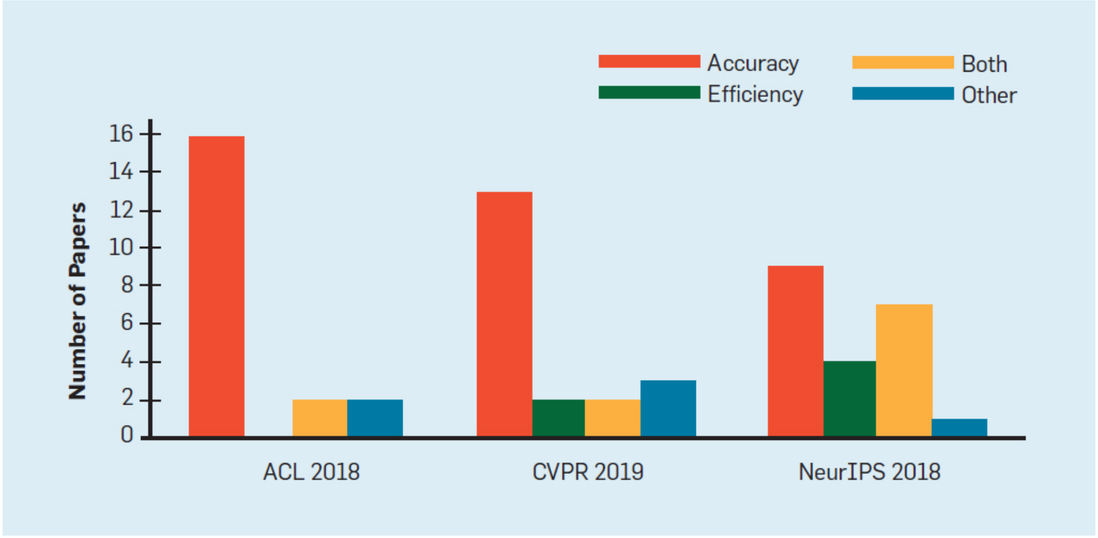
Efficient Methods

Mitigating the Challenges of Scaling

Enhanced Reporting

Efficient Methods

Accuracy or Efficiency?

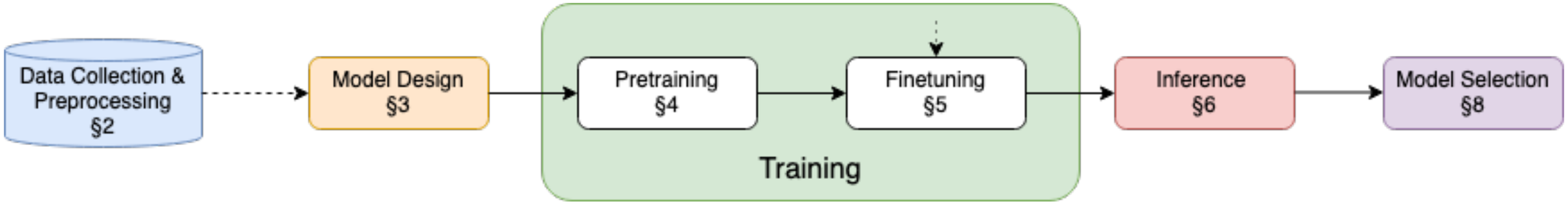


S. et al. (2020)

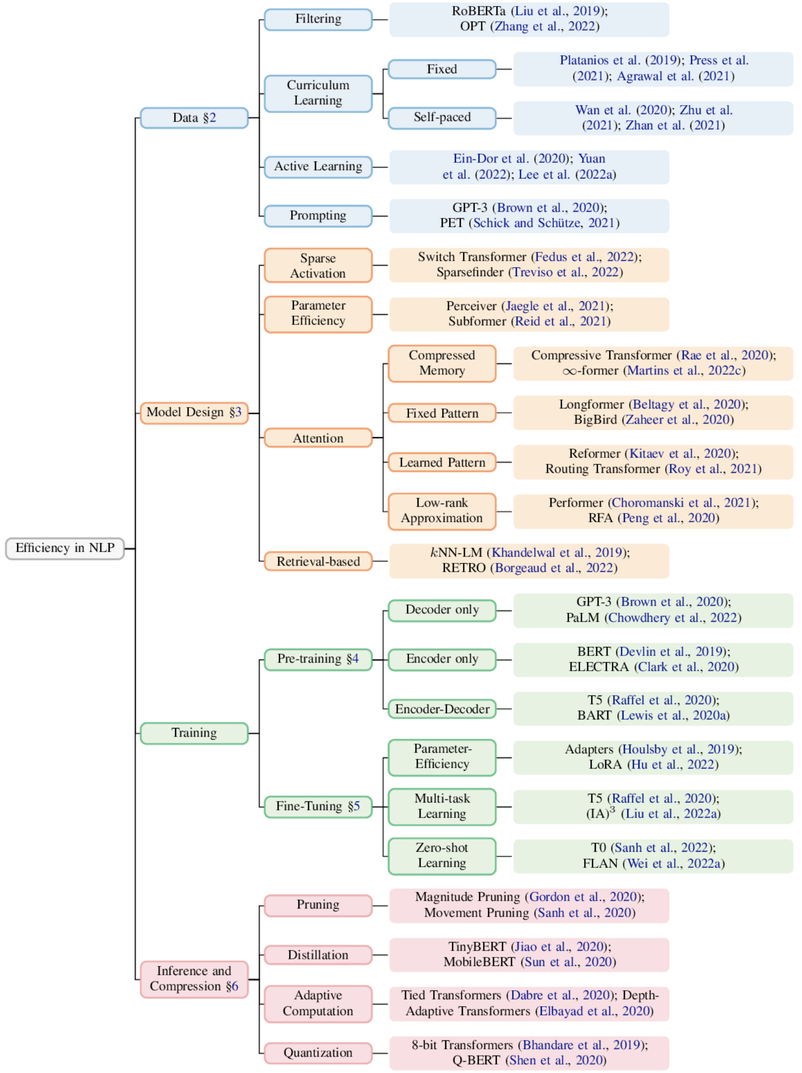
Efficient Methods for Natural Language Processing: A Survey

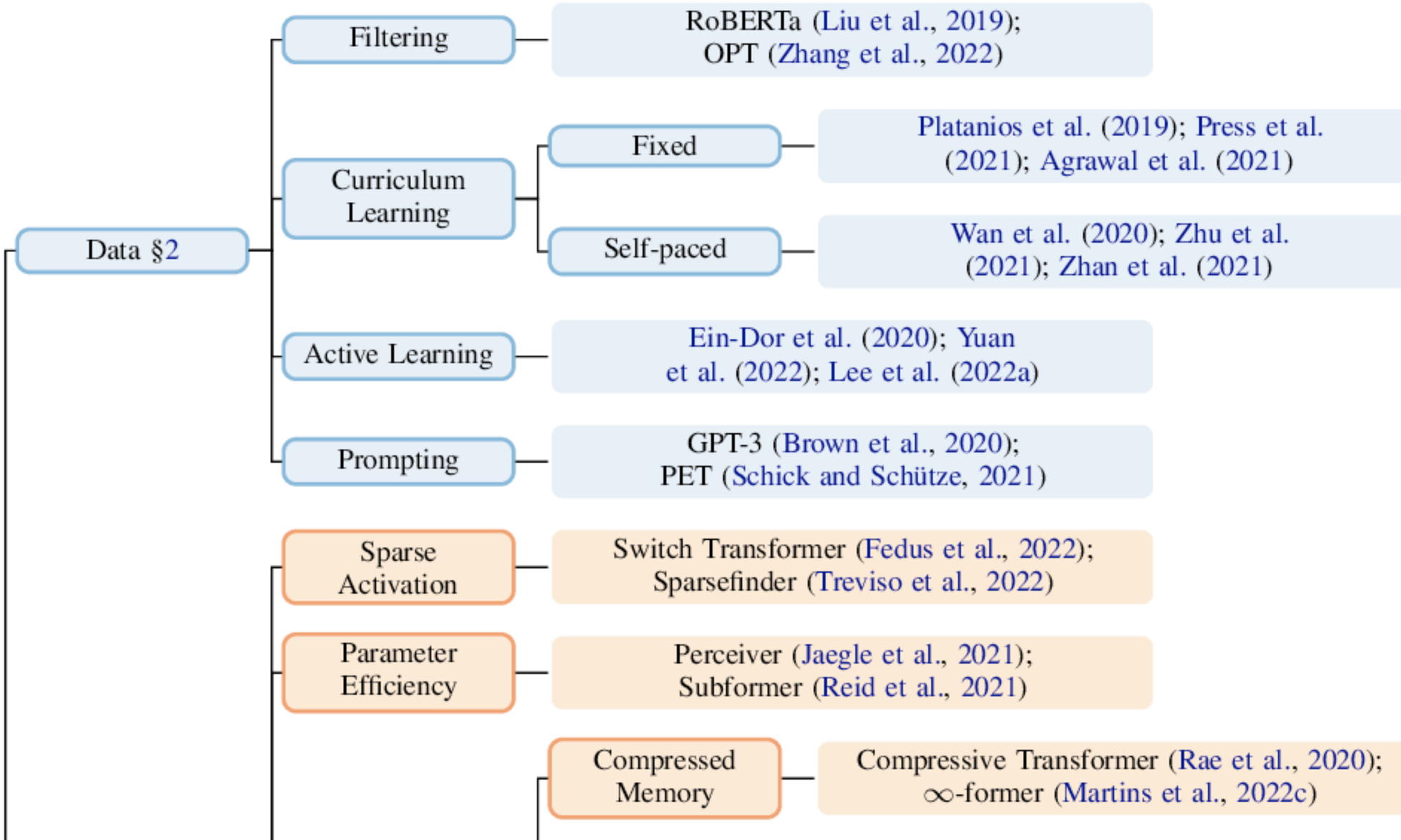
Marcos Treviso[10]*, Tianchu Ji[3]*, Ji-Ung Lee[7]*, Betty van Aken[8], Qingqing Cao[2], Manuel R. Ciosici[9], Michael Hassid[1], Kenneth Heafield[13], Sara Hooker[5], Pedro H. Martins[10], André F. T. Martins[10], Peter Milder[3], Colin Raffel[6], Edwin Simpson[4], Noam Slonim[12], Niranjan Balasubramanian[3], Leon Derczynski[11], Roy Schwartz[1]

[1]The Hebrew University of Jerusalem, [2]University of Washington, [3]Stony Brook University, [4]University of Bristol, [5]Cohere For AI, [6]University of North Carolina at Chapel Hill, [7]Technical University of Darmstadt, [8]Berliner Hochschule für Technik, [9]University of Southern California, [10]IST/University of Lisbon & Instituto de Telecomunicações, [11]IT University of Copenhagen, [12]IBM Research, [13]University of Edinburgh



Efficient Methods in NLP




Data §2
Filtering
RoBERTa (Liu et al., 2019); OPT (Zhang et al., 2022)
Curriculum Learning
Fixed
Platanios et al. (2019); Press et al. (2021); Agrawal et al. (2021)
Self-paced
Wan et al. (2020); Zhu et al. (2021); Zhan et al. (2021)
Active Learning
Ein-Dor et al. (2020); Yuan et al. (2022); Lee et al. (2022a)
Prompting
GPT-3 (Brown et al., 2020); PET (Schick and Schütze, 2021)
Sparse Activation
Switch Transformer (Fedus et al., 2022); Sparsefinder (Treviso et al., 2022)
Parameter Efficiency
Perceiver (Jaegle et al., 2021); Subformer (Reid et al., 2021)
Compressed Memory
Compressive Transformer (Rae et al., 2020); ∞-former (Martins et al., 2022c)

Filtering

- Non-text
  - Gibrish, HTML
- Text in other languages
- Foul text
- Typically done via simple, rule-based heuristics
  - Noisy process

Data-Efficient Training

Swayamdipta, S. et al., EMNLP 2020

- Not all training instances contribute the same to learning
  - Some are “*easy-to-learn*”, others are more challenging

Training Dynamics

- Training of neural networks typically requires iterating a large set of training examples

Training Dynamics

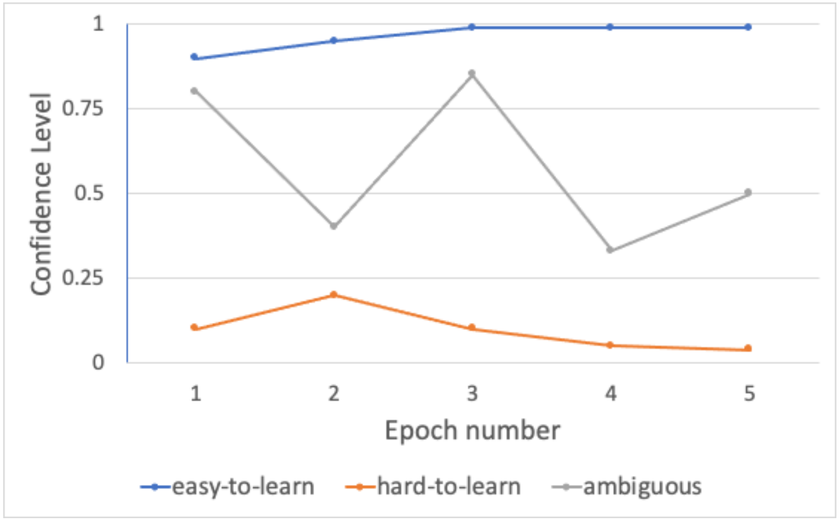
- Training of neural networks typically requires iterating a large set of training examples
- Models typically perform multiple iterations (aka epochs) over their training sets
  - In each of these iterations, the (partially-trained) model makes a prediction over each training instance

Training Dynamics

- Training of neural networks typically requires iterating a large set of training examples
- Models typically perform multiple iterations (aka epochs) over their training sets
  - In each of these iterations, the (partially-trained) model makes a prediction over each training instance
- These predictions provide a *valuable signal*

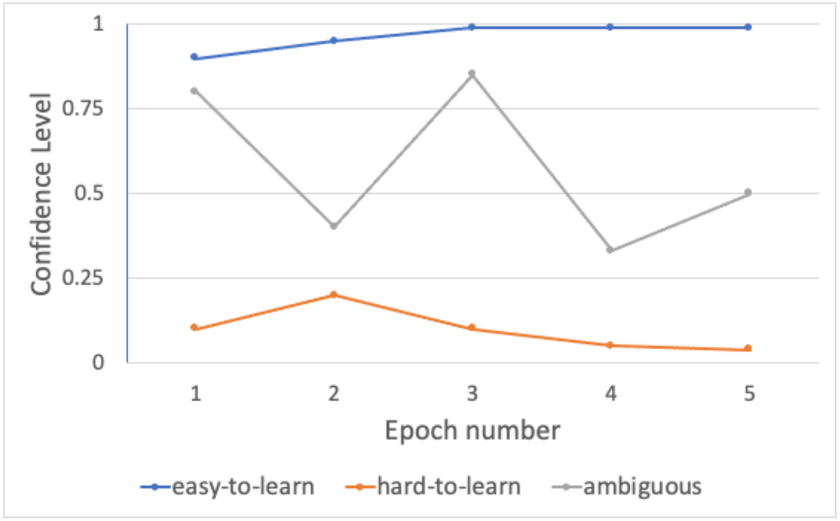
Training Dynamics

Toy Example



Training Dynamics

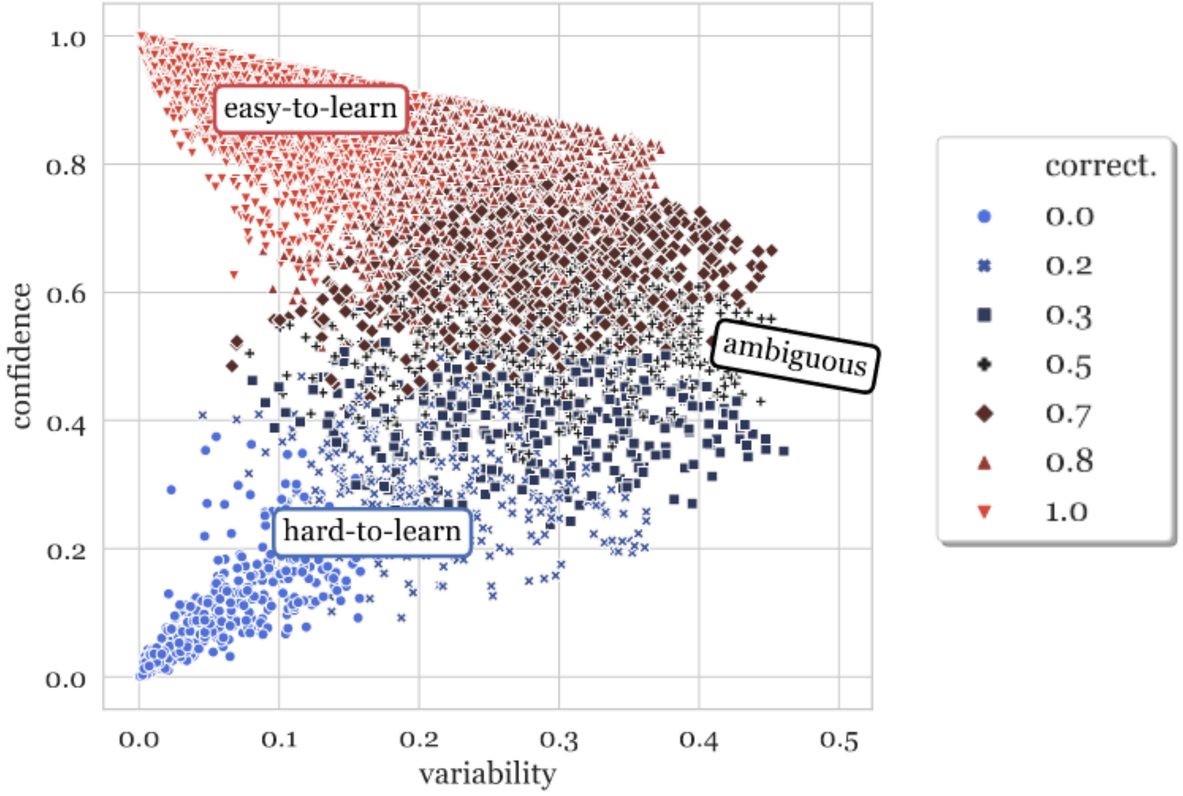
Toy Example



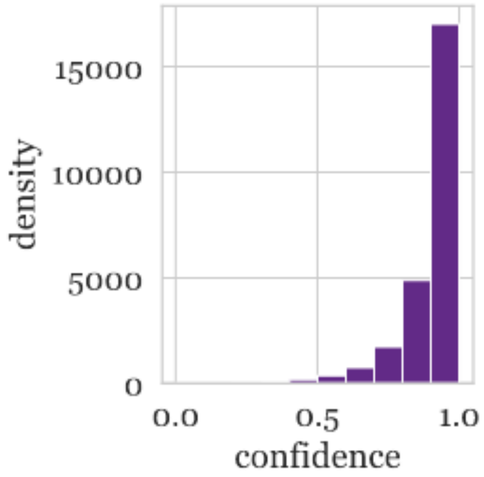
We extract two measures

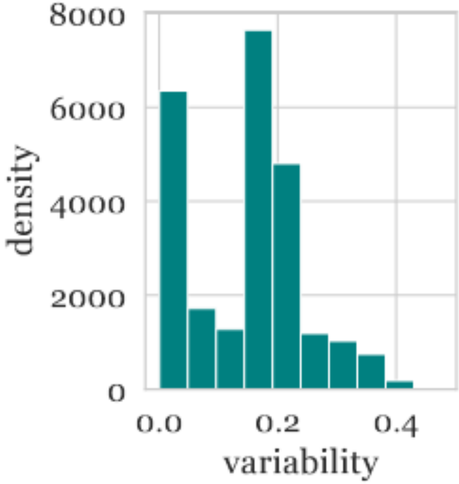
- Mean confidence
- Variability

Dataset Map


easy-to-learn
ambiguous
hard-to-learn
confidence
variability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
correct.
0.0
0.2
0.3
0.5
0.7
0.8
1.0

Most Instances are Easy-to-Learn





Examples

	Instance	Option1	Option2
easy-to-learn	The man chose to buy the roses instead of the carnations because the __ were more beautiful.	roses*	carnations
	We enjoyed the meeting tonight but not the play as the __ was rather dull.	meeting	play*
ambiguous	The dog ran up to Leslie and away from Lawrence because __ had soap for the dog to take a bath.	Leslie^{-}	Lawrence*
	Kayla dated many more people at once than Betty, because __ was in an exclusive relationship.	Kayla*	Betty^{+}

Examples

	Instance	Option1	Option2
easy-to-learn	The man chose to buy the roses instead of the carnations because the __ were more beautiful.	roses*	carnations
	We enjoyed the meeting tonight but not the play as the __ was rather dull.	meeting	play*
ambiguous	The dog ran up to Leslie and away from Lawrence because __ had soap for the dog to take a bath.	Leslie^{-}	Lawrence*
	Kayla dated many more people at once than Betty, because __ was in an exclusive relationship.	Kayla*	Betty^{+}

Examples

	Instance	Option1	Option2
easy-to-learn	The man chose to buy the roses instead of the carnations because the __ were more beautiful.	roses*	carnations
	We enjoyed the meeting tonight but not the play as the __ was rather dull.	meeting	play*
ambiguous	The dog ran up to Leslie and away from Lawrence because __ had soap for the dog to take a bath.	Leslie^{-}	Lawrence*
	Kayla dated many more people at once than Betty, because __ was in an exclusive relationship.	Kayla*	Betty^{+}

Sample-Efficient Training

- *easy-to-learn* instances provide little value to training
- Can we use training dynamics to select the *most valuable* instances?

Experiments

WinoGrande, RoBERTa-Large

	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

33%

	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

33%

	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

33%

	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

33%

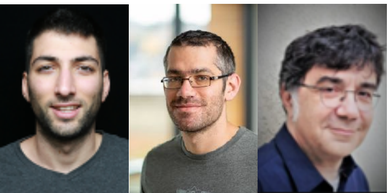
	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Data Efficient Masked Language Modeling for Vision and Language

Bitton, Stanovsky, Elhadad & S., Findings of EMNLP 2021

- Given an image and a caption describing it
  - Mask some of the words in the caption
  - Train model to complete the masked words

A tiger [MASK] eating the carrot

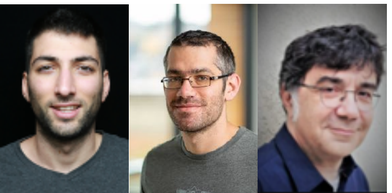


Data Efficient Masked Language Modeling for Vision and Language

Bitton, Stanovsky, Elhadad & S., Findings of EMNLP 2021

- Given an image and a caption describing it
  - Mask some of the words in the caption
  - Train model to complete the masked words
- Words are randomly masked with a 15% probability

A tiger [MASK] eating the carrot



Problems with Pre-training Vision and Language Models

- Of the masked tokens, roughly one half are *stop-words* or *punctuation*

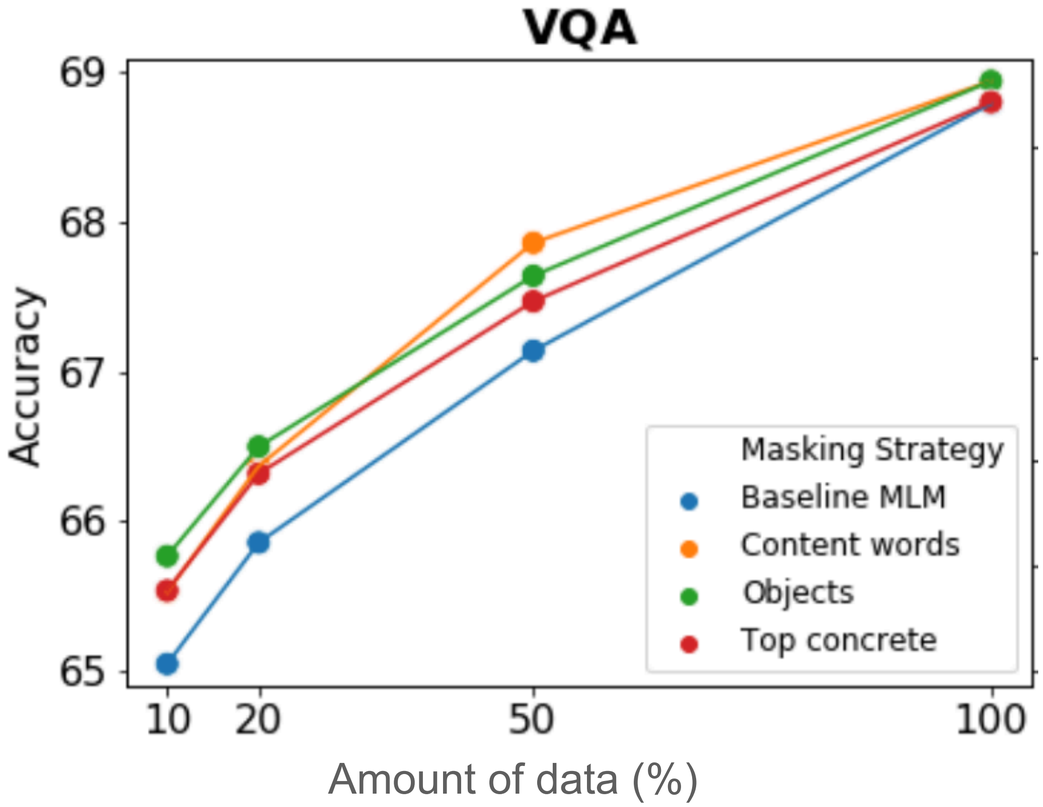
Problems with Pre-training Vision and Language Models

- Of the masked tokens, roughly one half are *stop-words* or *punctuation*
- Image captions are typically short (6-7 words)
  - ⇒ in roughly 1/3 of the captions, **no word is masked**!

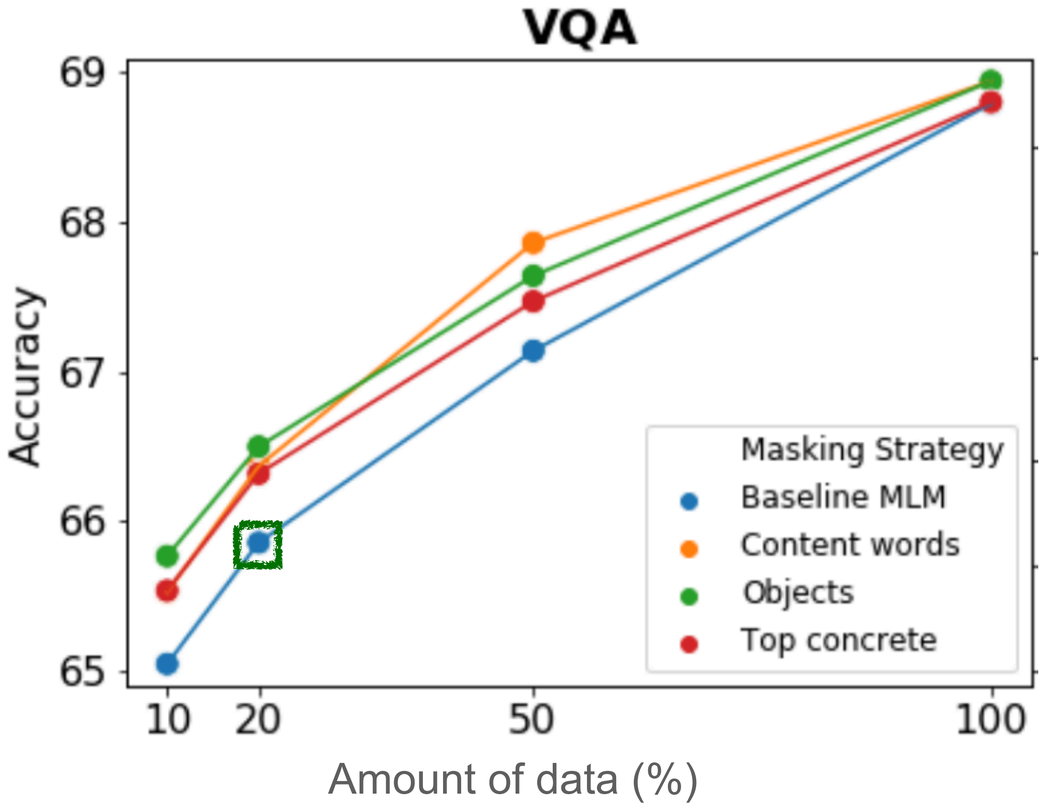
Problems with Pre-training Vision and Language Models

- Of the masked tokens, roughly one half are *stop-words* or *punctuation*
- Image captions are typically short (6-7 words)
  - ⇒ in roughly 1/3 of the captions, **no word is masked**!
- **Improved** pre-training:
  - Mask **exactly one token** in each document
  - Only mask **content words**

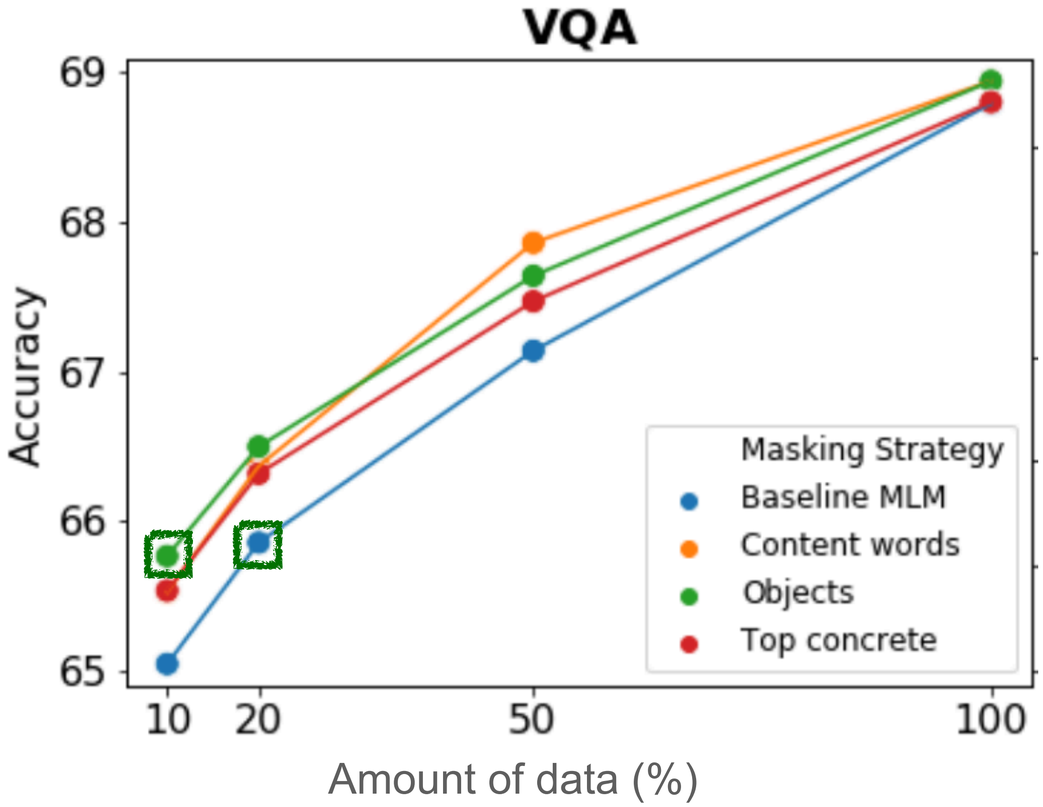
Data Efficient Training



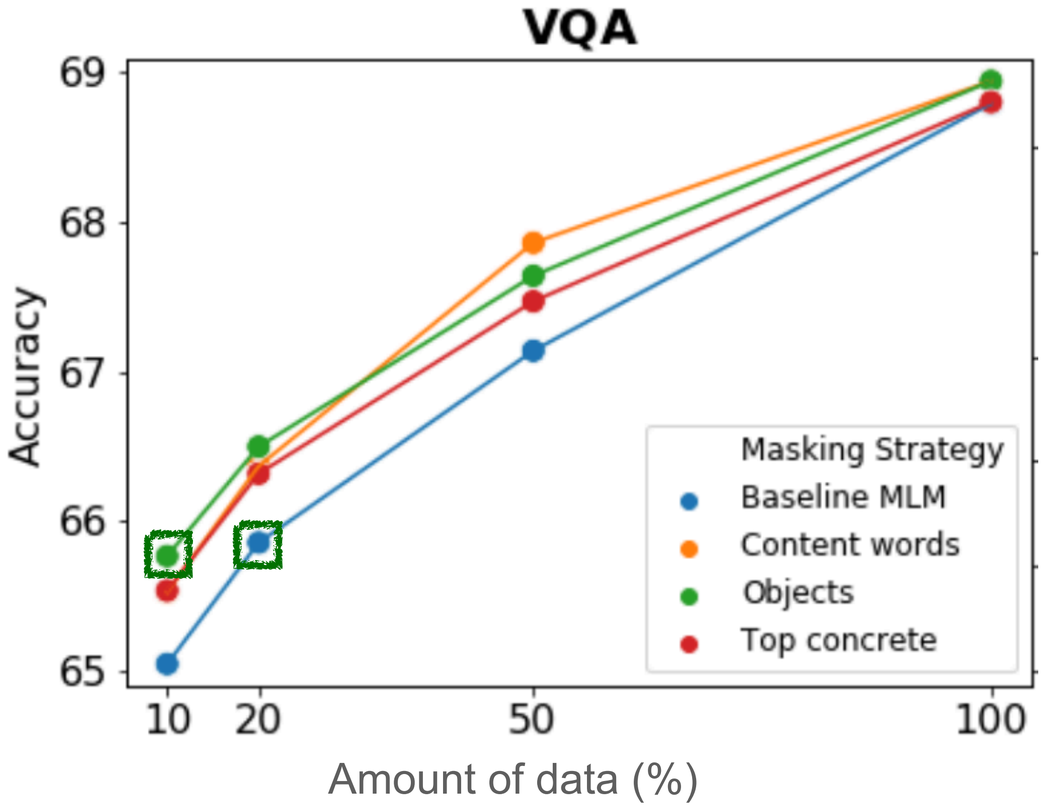
Data Efficient Training



Data Efficient Training



Data Efficient Training



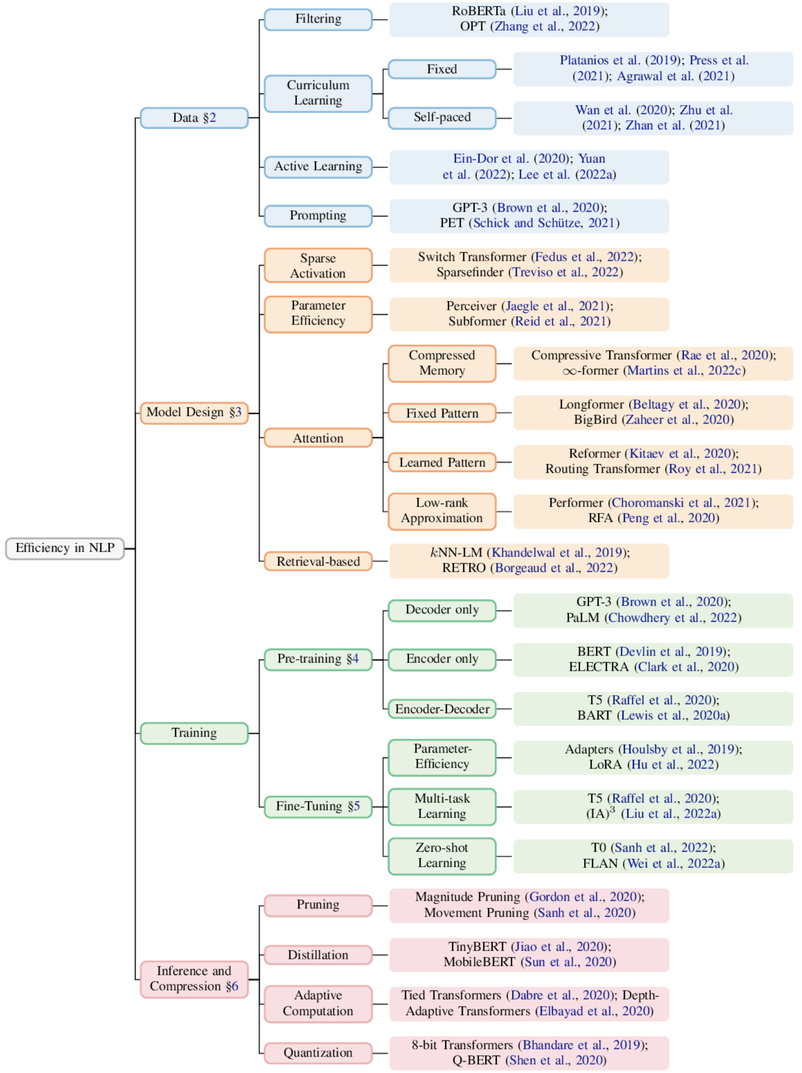
Similar accuracy, twice as fast

Data Efficiency

Open Questions

- Do we really need massive web-scale data to train our models?
  - Can we get along with less?

Efficient Methods in NLP

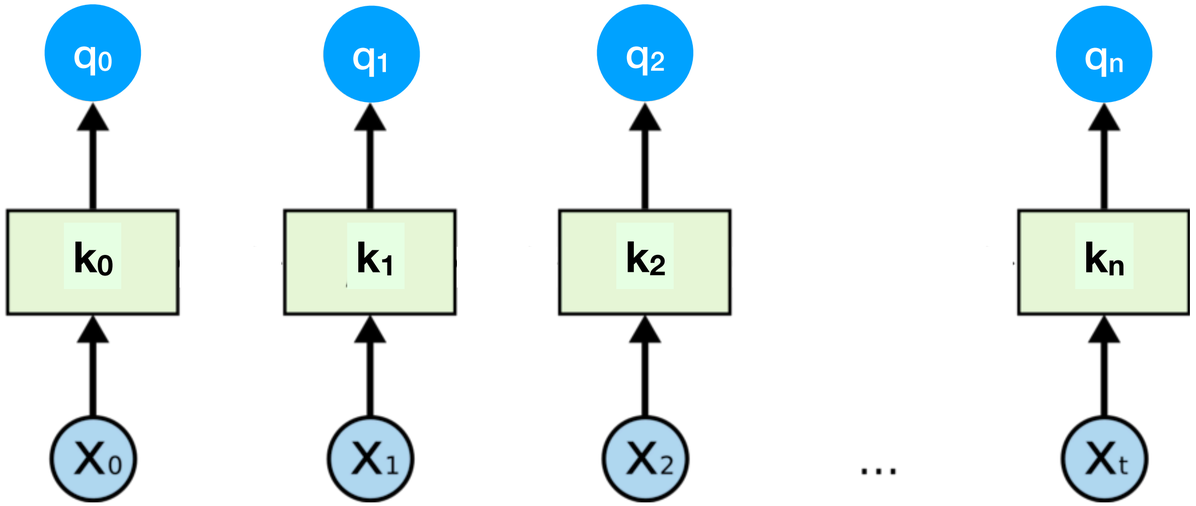


- Efficiency in NLP
  - Model Design §3
    - Active Learning — Ein-Dor et al. (2020); Yuan et al. (2022); Lee et al. (2022a)
    - Prompting — GPT-3 (Brown et al., 2020); PET (Schick and Schütze, 2021)
    - Sparse Activation — Switch Transformer (Fedus et al., 2022); Sparsefinder (Treviso et al., 2022)
    - Parameter Efficiency — Perceiver (Jaegle et al., 2021); Subformer (Reid et al., 2021)
    - Attention
      - Compressed Memory — Compressive Transformer (Rae et al., ∞-former (Martins et al., 2022c
      - Fixed Pattern — Longformer (Beltagy et al., 2020 BigBird (Zaheer et al., 2020)
      - Learned Pattern — Reformer (Kitaev et al., 2020); Routing Transformer (Roy et al., 2
      - Low-rank Approximation — Performer (Choromanski et al., 202 RFA (Peng et al., 2020)
    - Retrieval-based — kNN-LM (Khandelwal et al., 2019); RETRO (Borgeaud et al., 2022)
    - Decoder only — GPT-3 (Brown et al., 2020); PaLM (Chowdhery et al., 2022)

Transformers

Vaswani et al., 2017

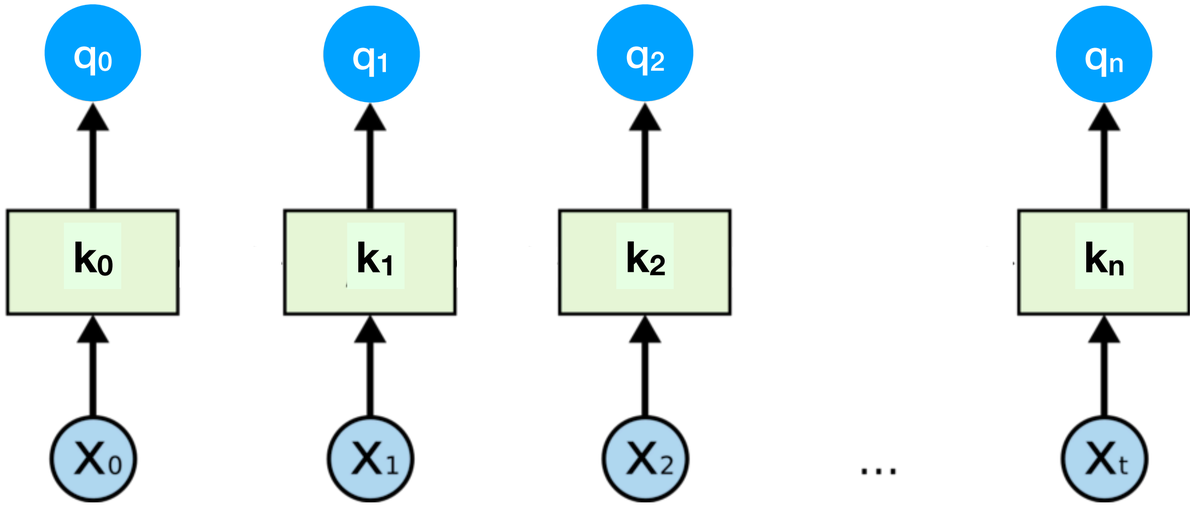
- **The** method for text representation
  - Also for vision, speech, combio, …



Transformers

Vaswani et al., 2017

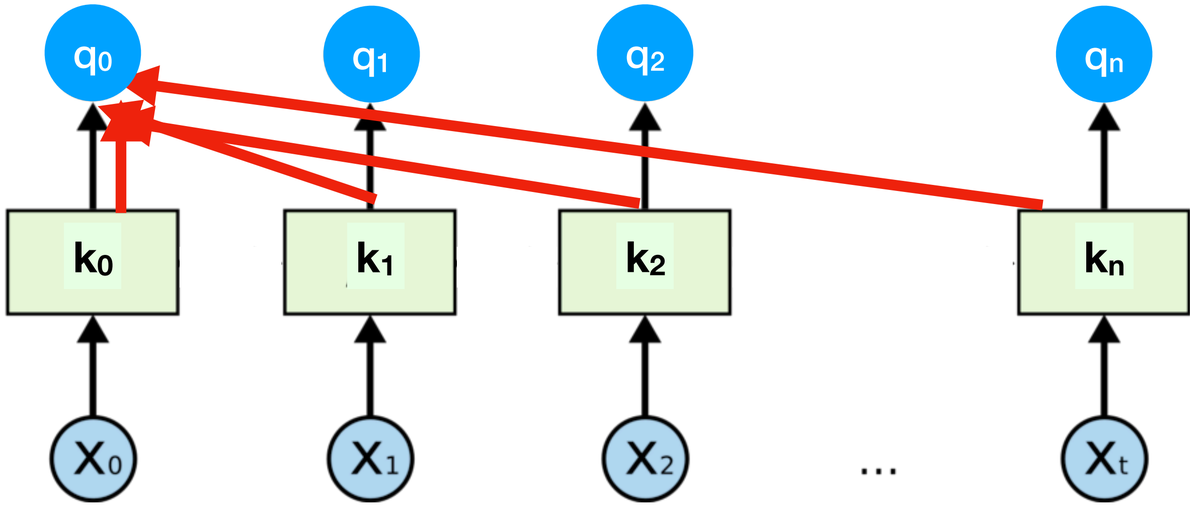
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

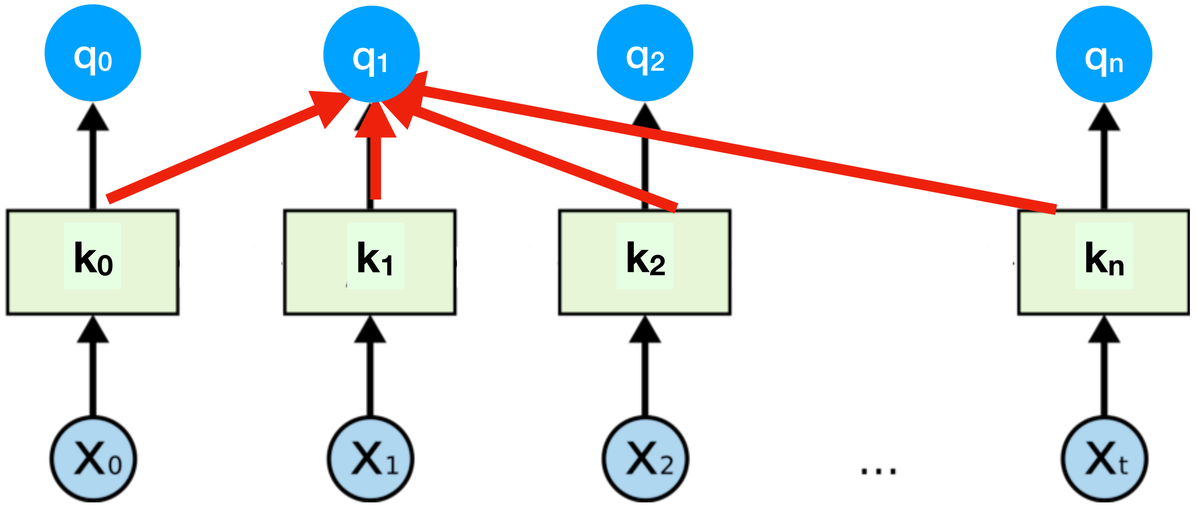
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

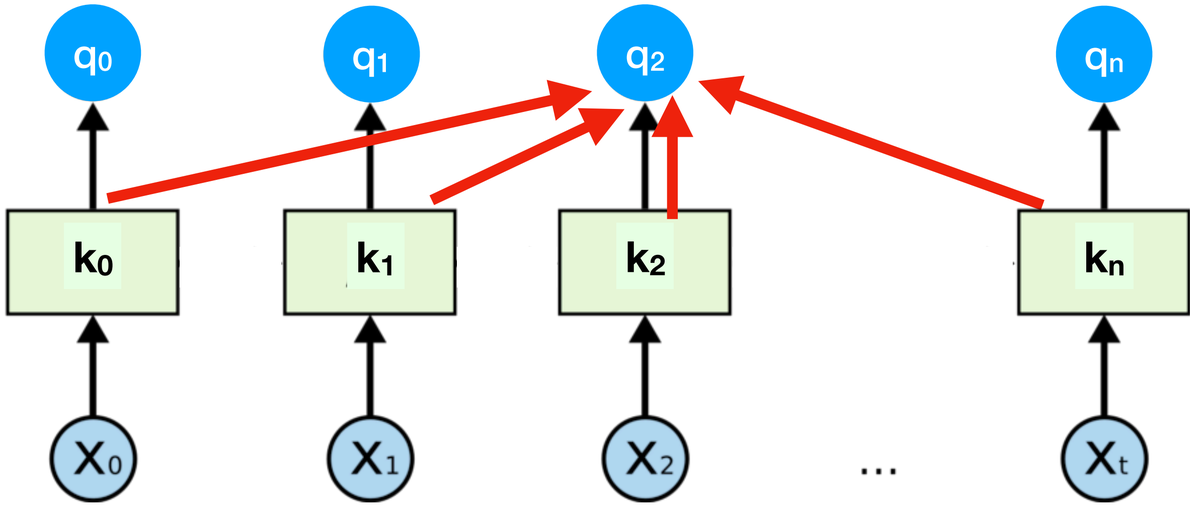
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

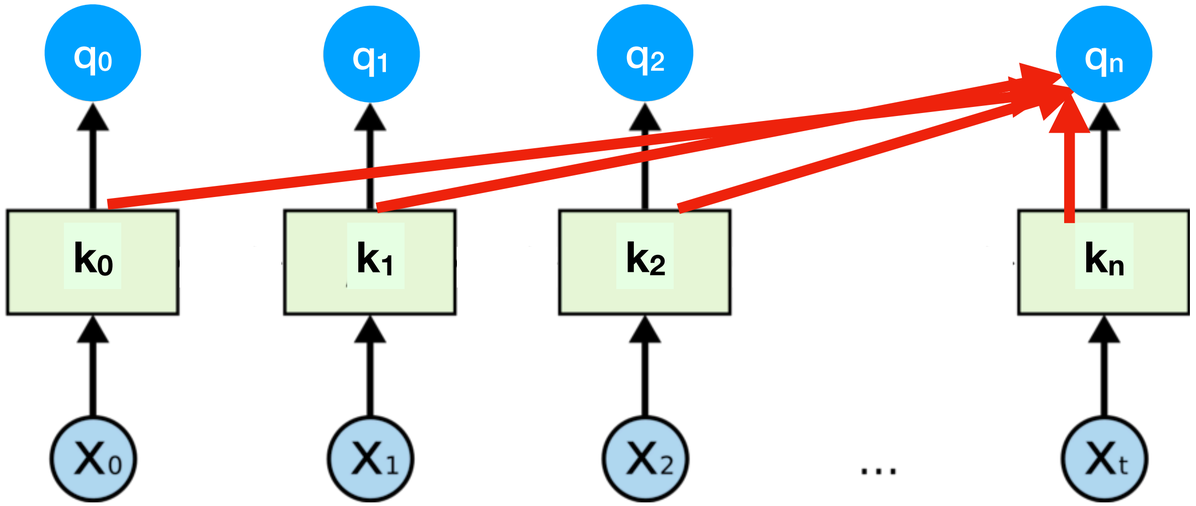
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

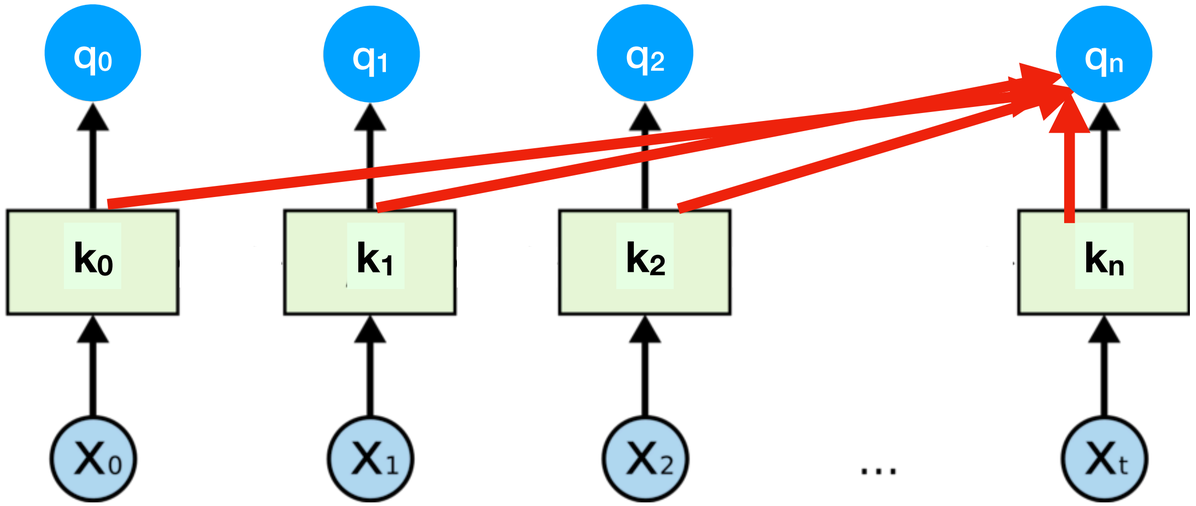
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

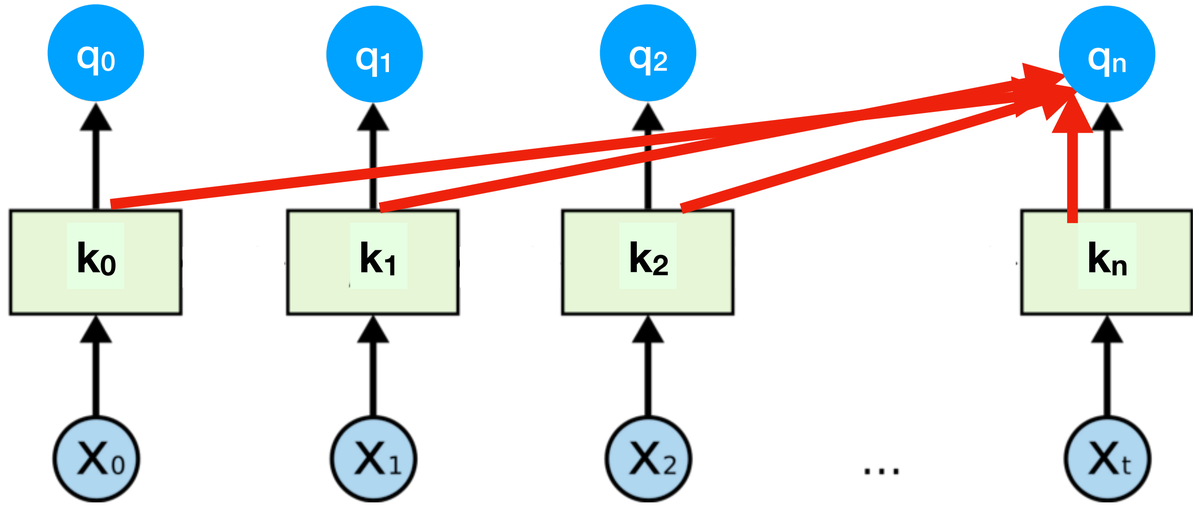
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words
- O(n^2) complexity in the sentence length n



Transformers

Vaswani et al., 2017

- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words
- O(n^2) complexity in the sentence length n
- Fatal for long sequences
  - Books, articles, etc.

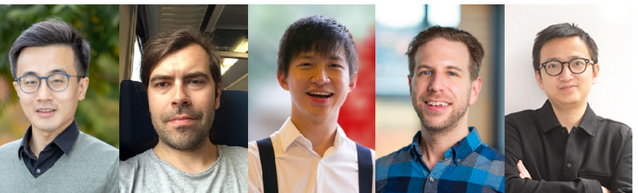


Random Feature Attention

Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

- **Key idea**: approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)

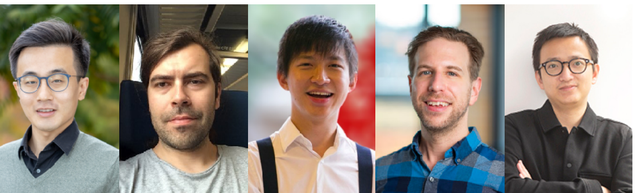


Random Feature Attention

Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

- **Key idea**: approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math

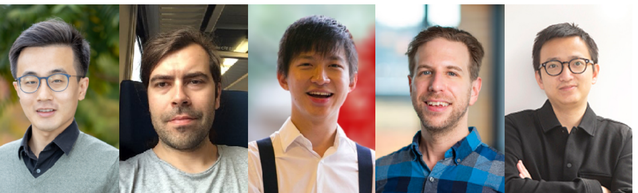


Random Feature Attention

Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

- **Key idea**: approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math

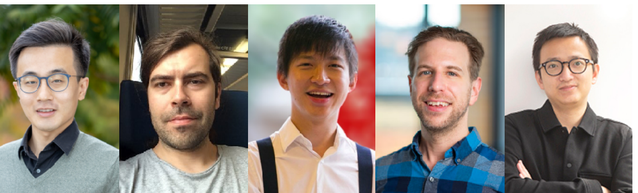


Random Feature Attention

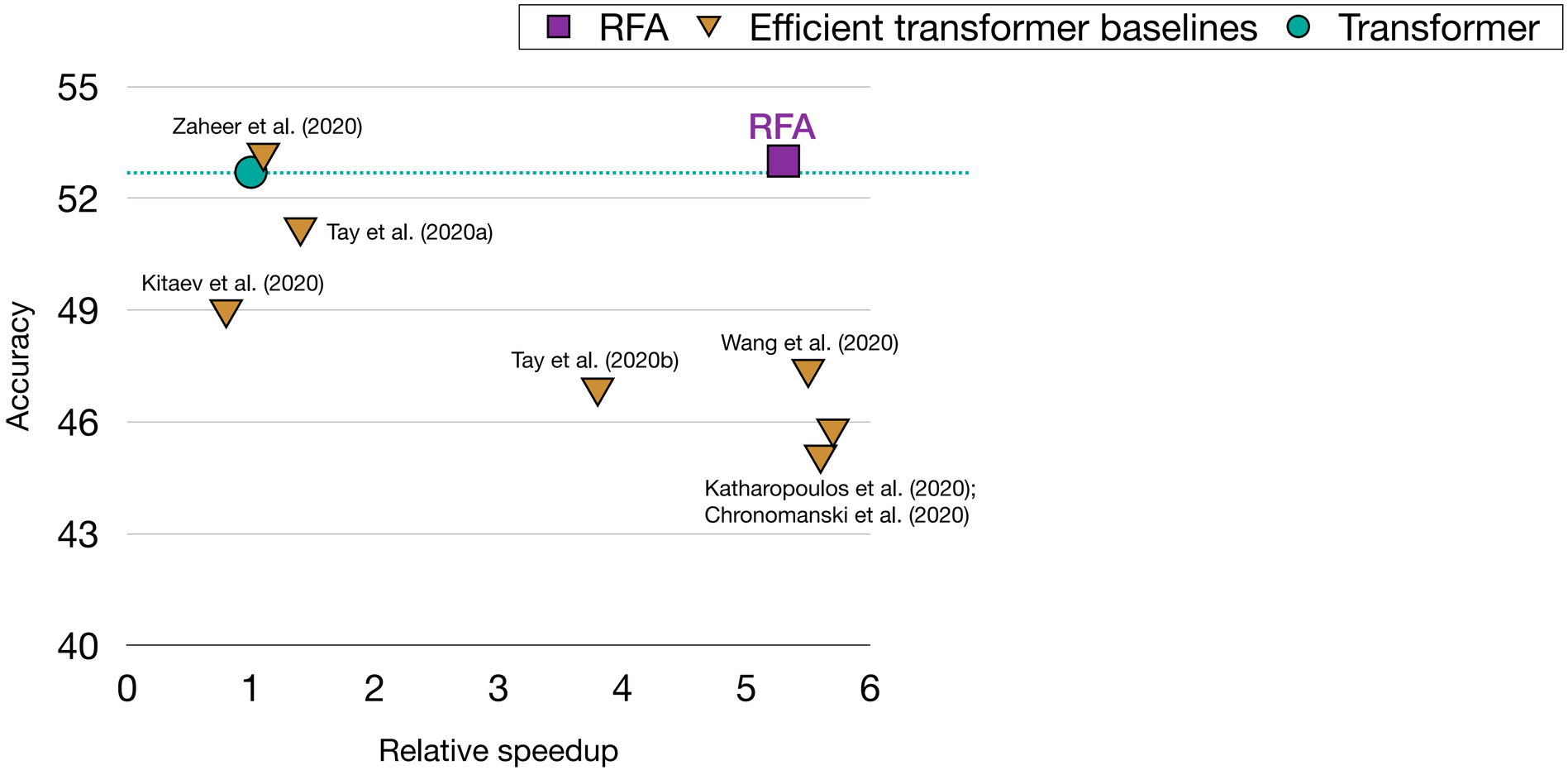
Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

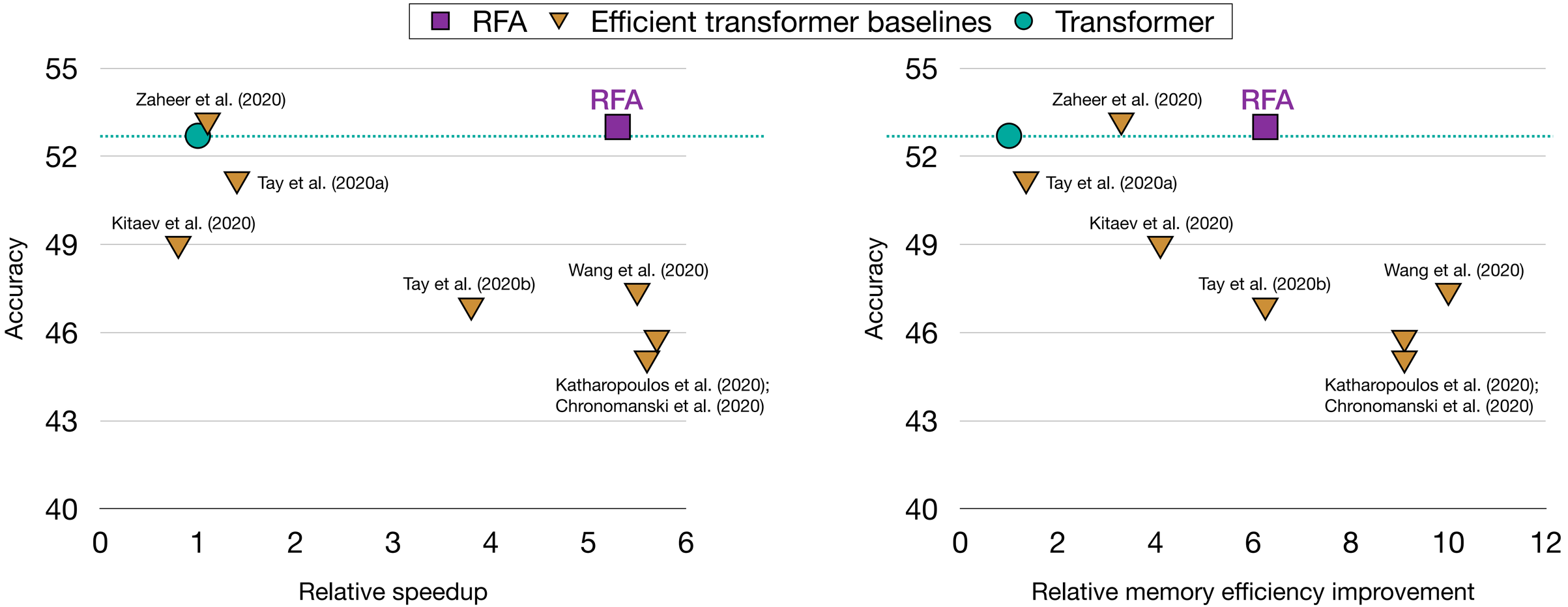
- **Key idea**: approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math
- Linear runtime and memory requirements



Better Efficiency-Accuracy Tradeoff



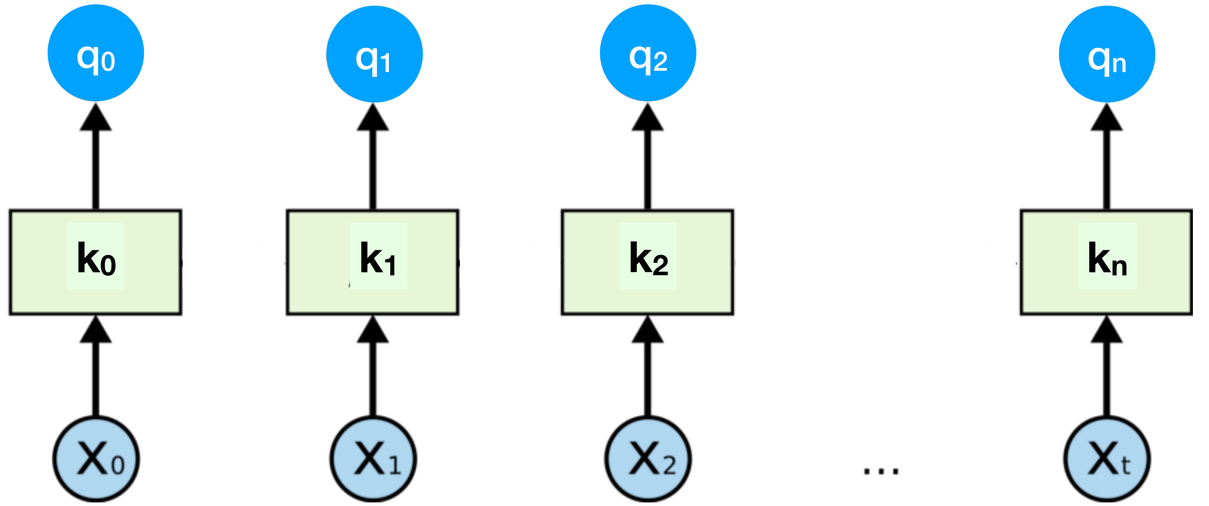
Better Efficiency-Accuracy Tradeoff



ABC: Attention with Bounded-memory Control

Peng, Kasai, Pappas, Yogatama, Wu, Kong, **S.** & Smith, ACL 2022

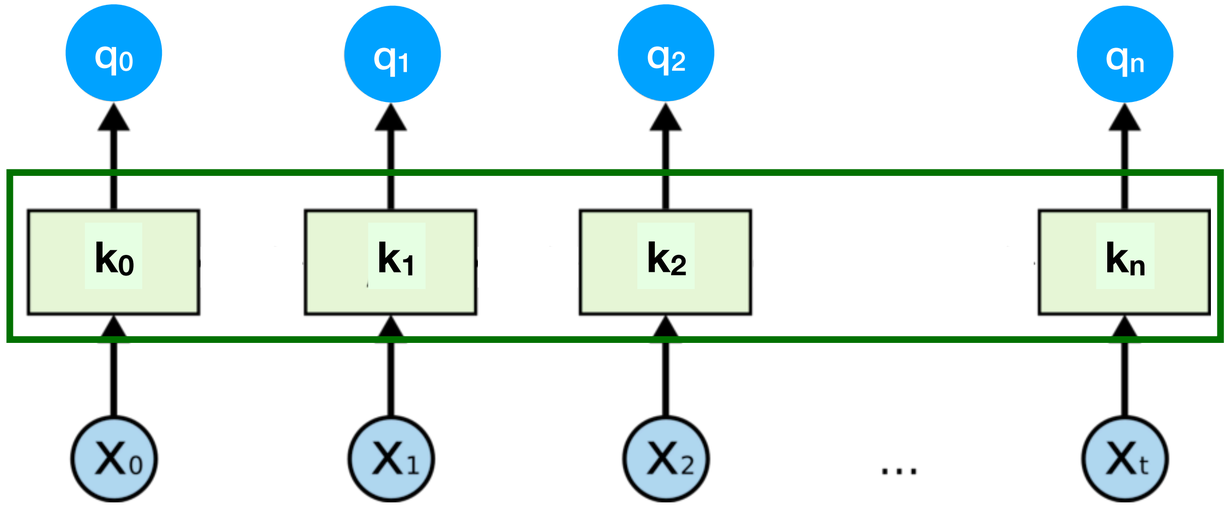
- **Key intuition**: treat the sentence as **memory of size *n***



ABC: Attention with Bounded-memory Control

Peng, Kasai, Pappas, Yogatama, Wu, Kong, **S.** & Smith, ACL 2022

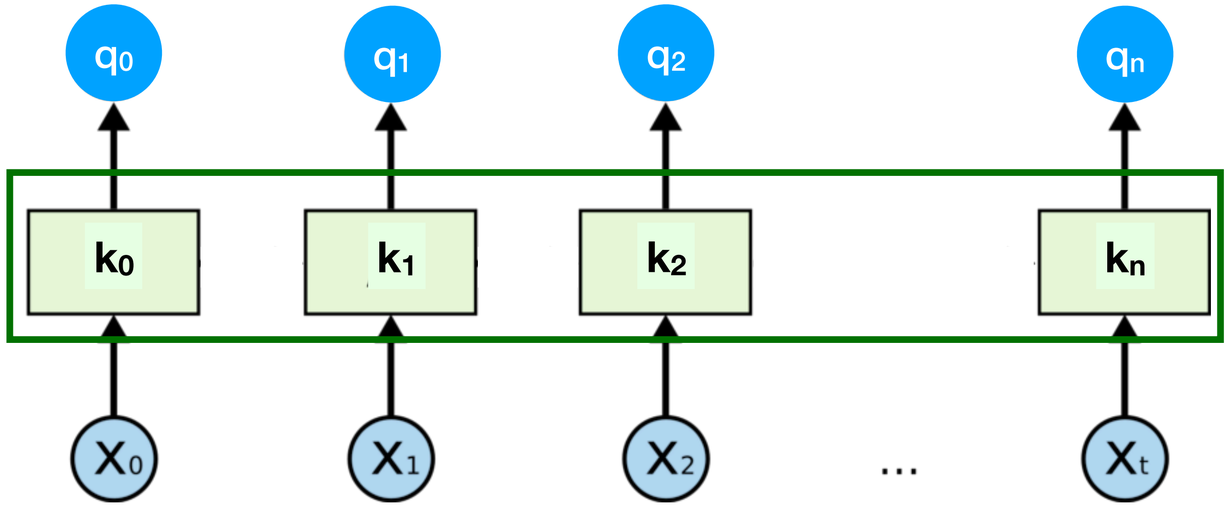
- **Key intuition**: treat the sentence as **memory of size *n***



ABC: Attention with Bounded-memory Control

Peng, Kasai, Pappas, Yogatama, Wu, Kong, S. & Smith, ACL 2022

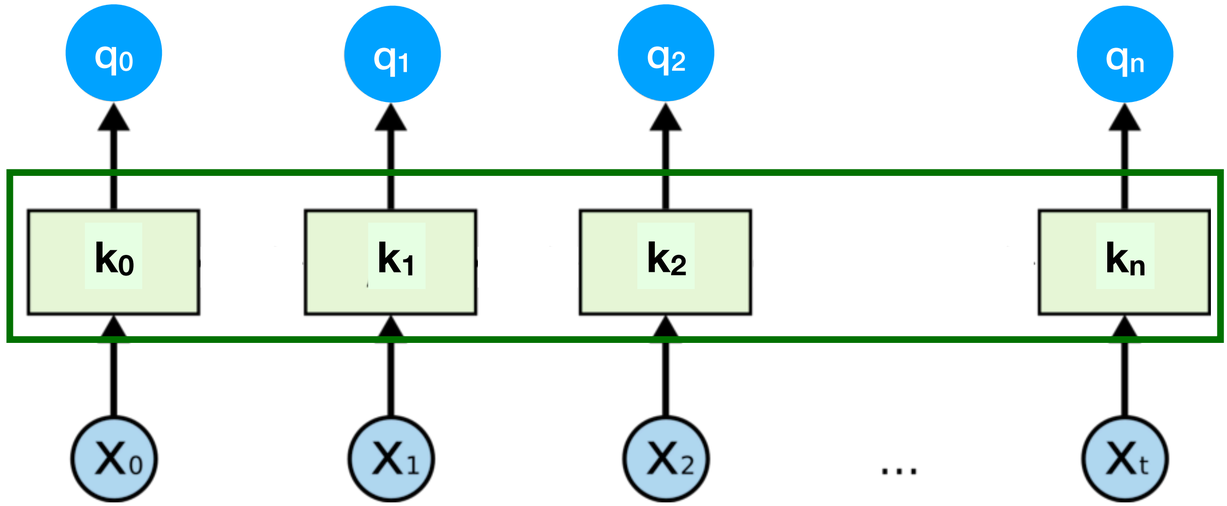
- **Key intuition**: treat the sentence as **memory of size *n***
- **Key idea**: replace this memory with a fixed size memory of (fixed) size *k* << *n*
  - Instead of attending *n* tokens, each word attends to *k* tokens



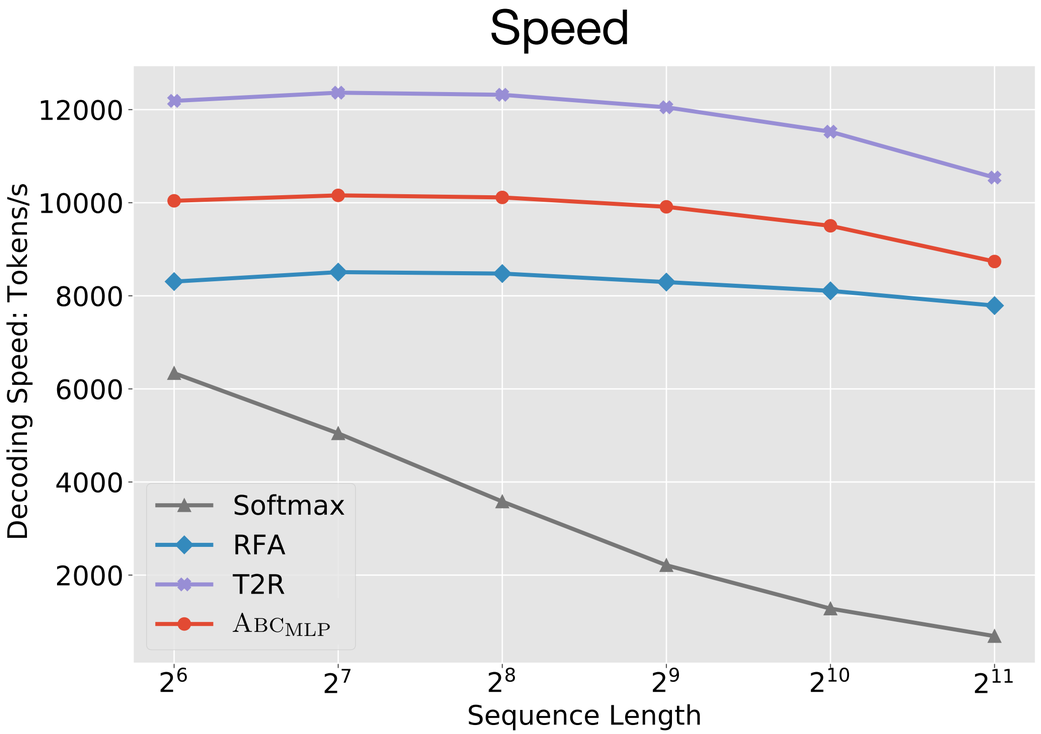
ABC: Attention with Bounded-memory Control

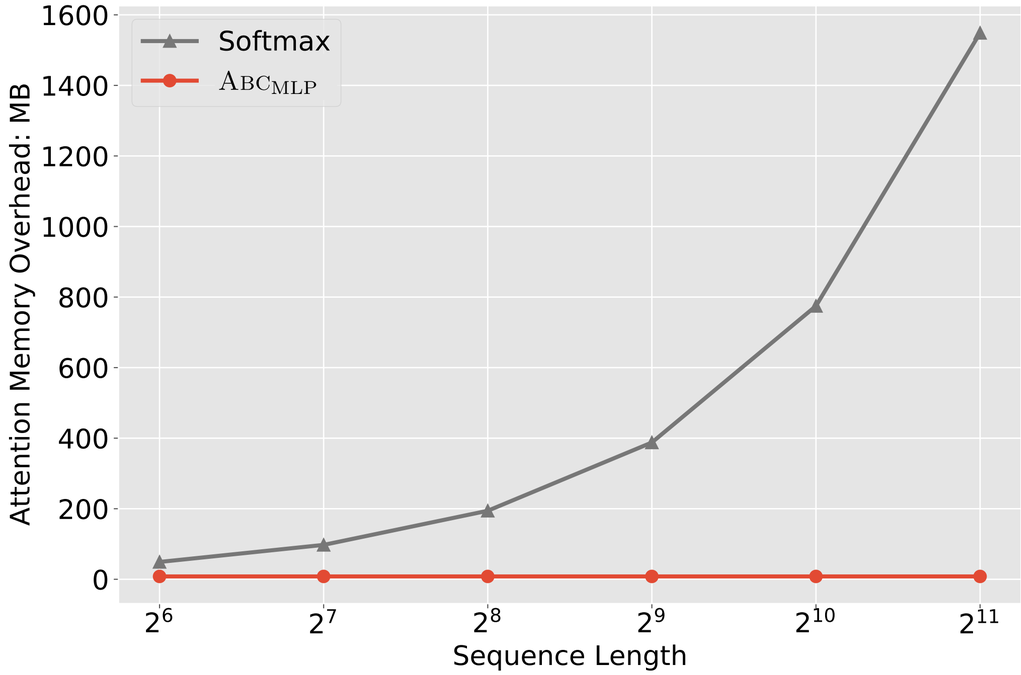
Peng, Kasai, Pappas, Yogatama, Wu, Kong, S. & Smith, ACL 2022

- **Key intuition**: treat the sentence as **memory of size *n***
- **Key idea**: replace this memory with a fixed size memory of (fixed) size *k* << *n*
  - Instead of attending *n* tokens, each word attends to *k* tokens
- Overall complexity linear in *n*
  - With constant *k*



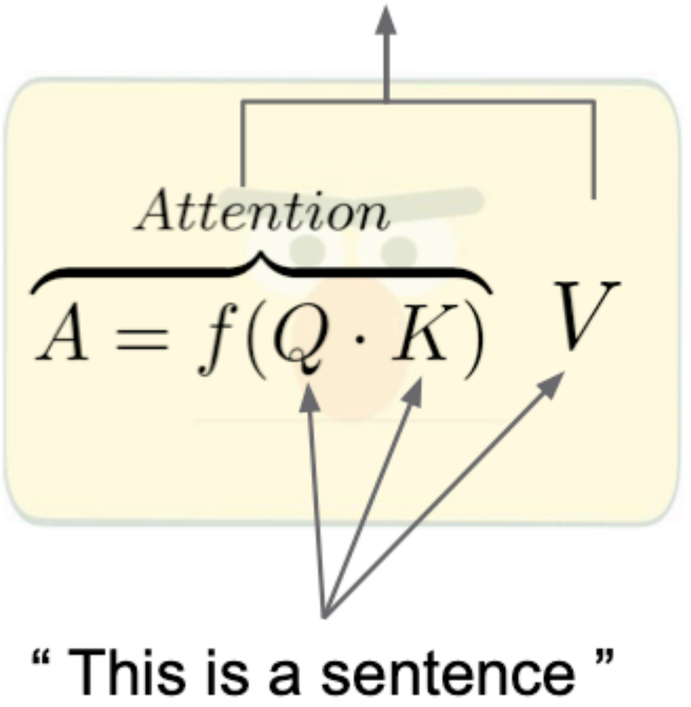
ABC Results

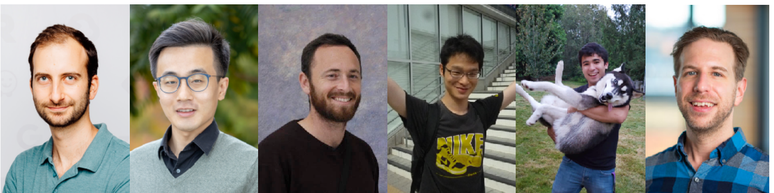

Speed
Decoding Speed: Tokens/s
12000
10000
8000
6000
4000
2000
Softmax
RFA
T2R
ABC_MLP
2^6
2^7
2^8
2^9
2^10
2^11
Sequence Length


Attention Memory Overhead: MB
1600
1400
1200
1000
800
600
400
200
0
Softmax
ABC_MLP
2^6
2^7
2^8
2^9
2^10
2^11
Sequence Length

How Much Does Attention Actually Attend?

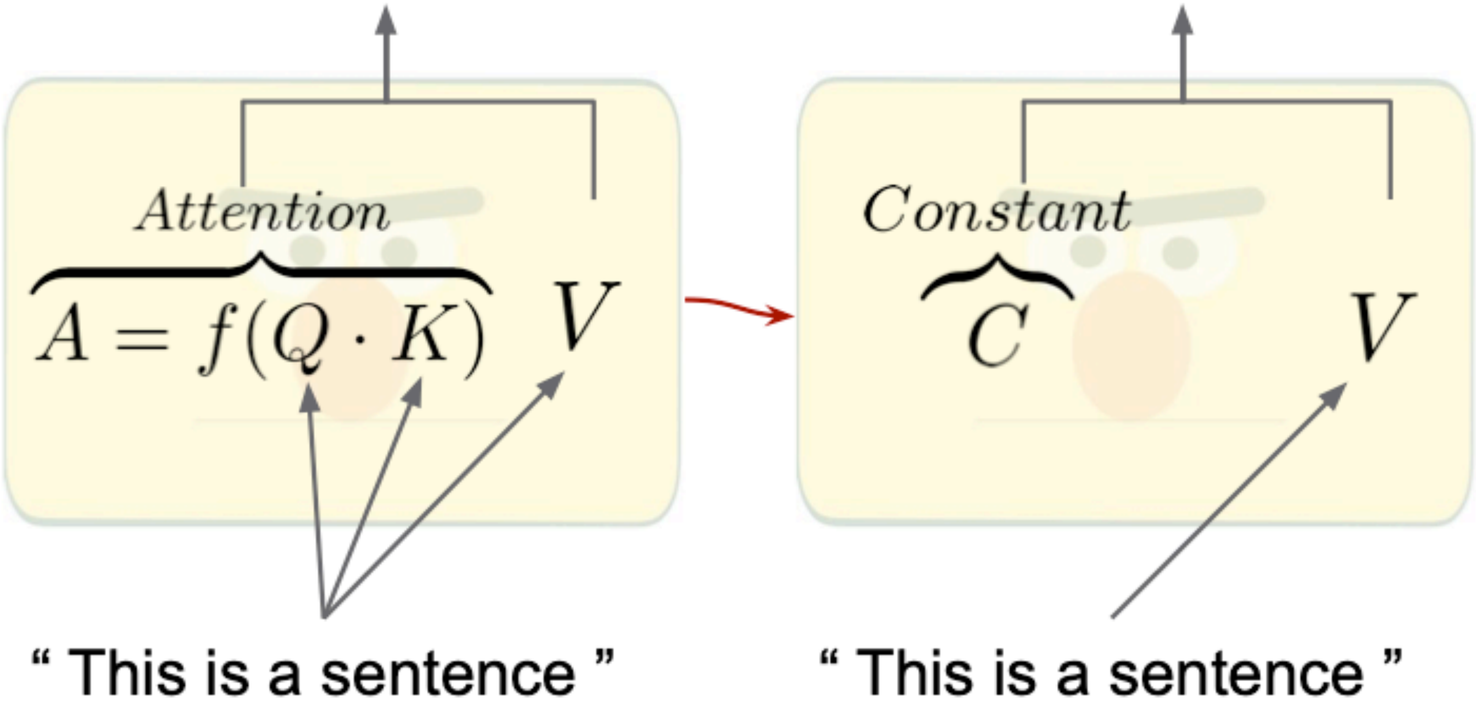
Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

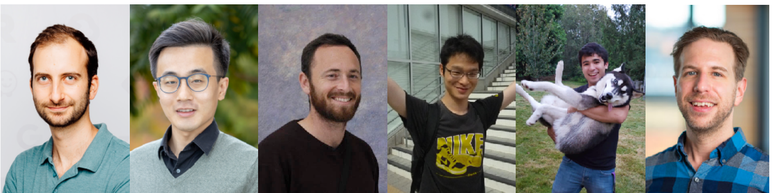




How Much Does Attention Actually Attend?

Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

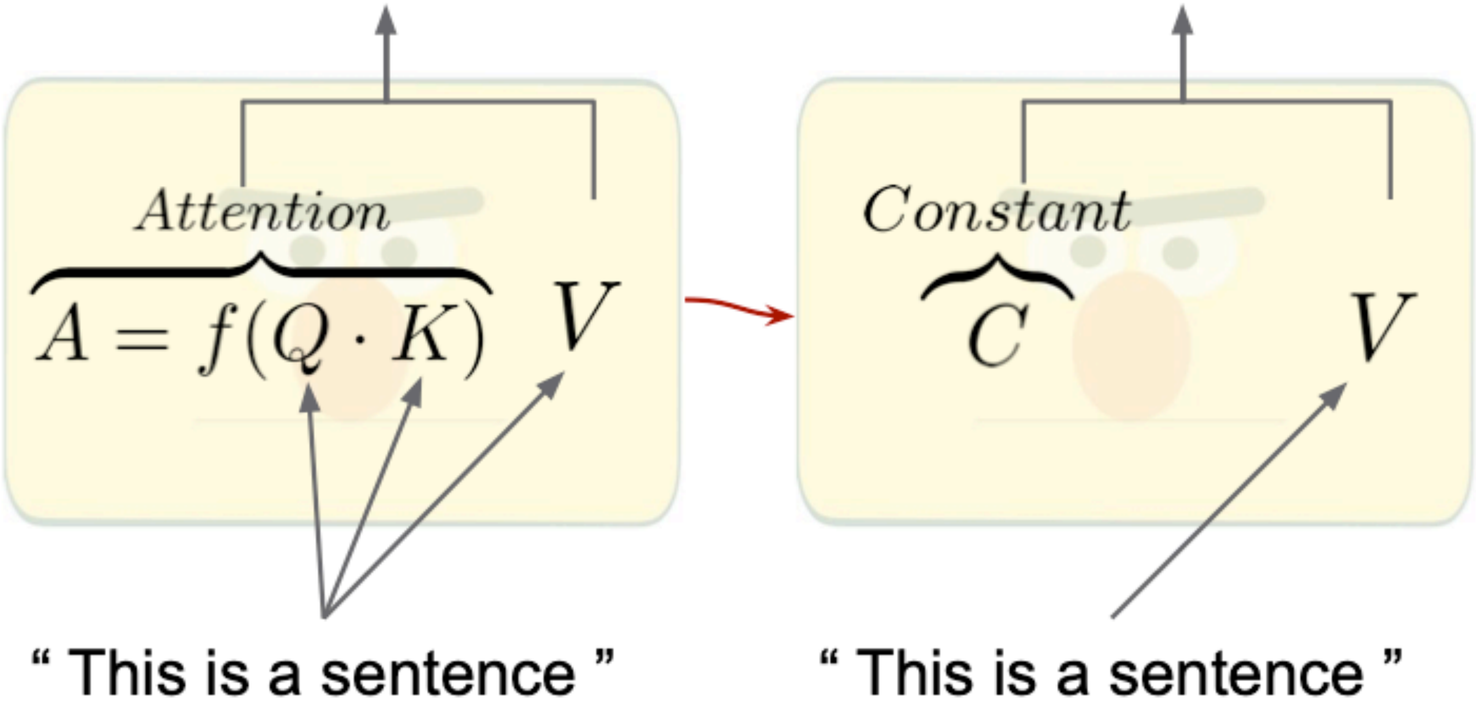


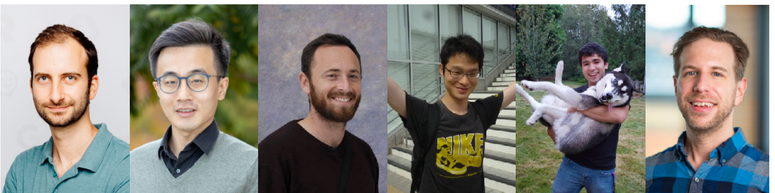


How Much Does Attention Actually Attend?

Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

- Model doesn't collapse
  - Average accuracy loss of 8% only

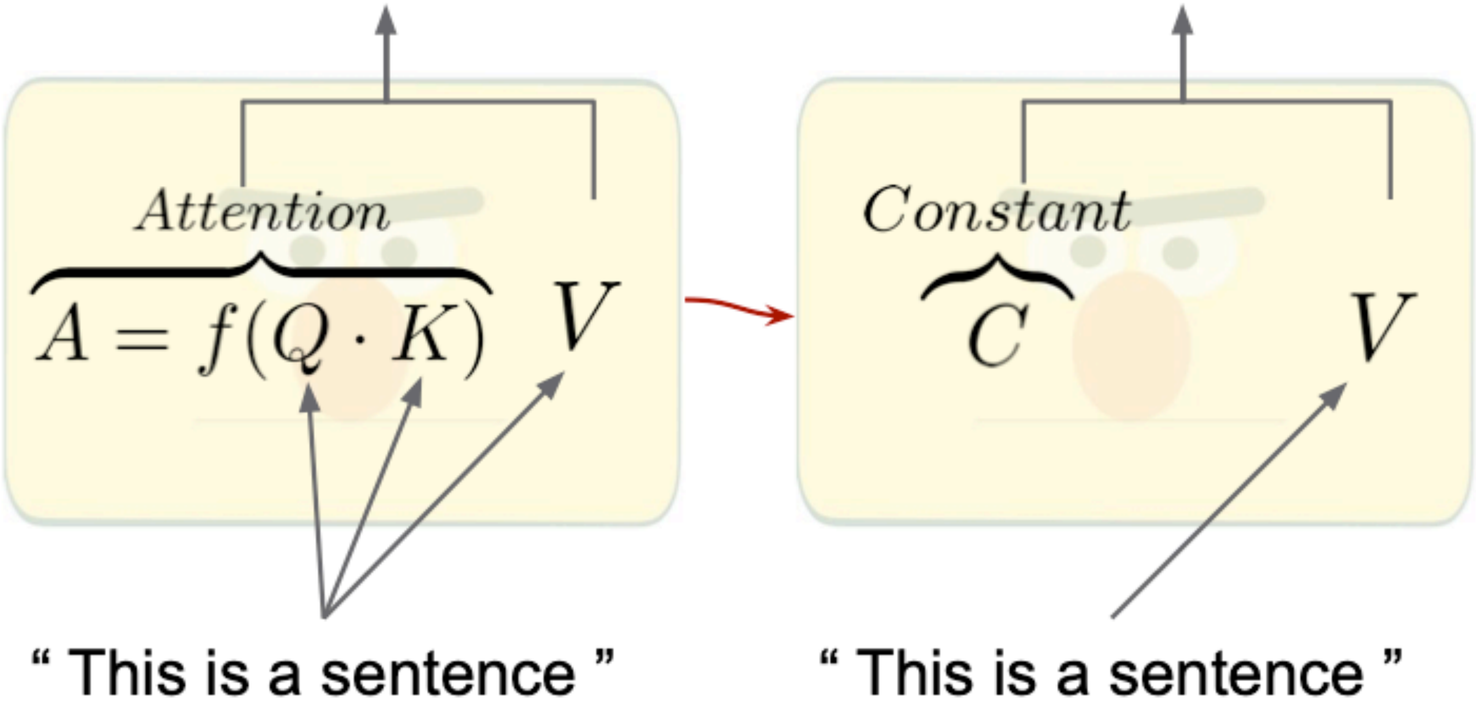


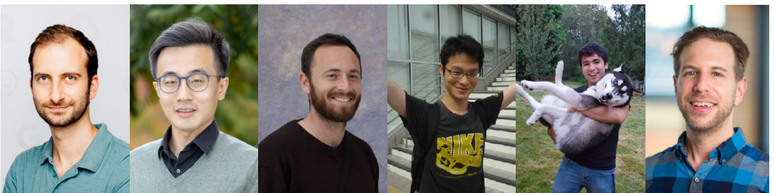


How Much Does Attention Actually Attend?

Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

- Model doesn't collapse
  - Average accuracy loss of 8% only
- Potential for huge savings



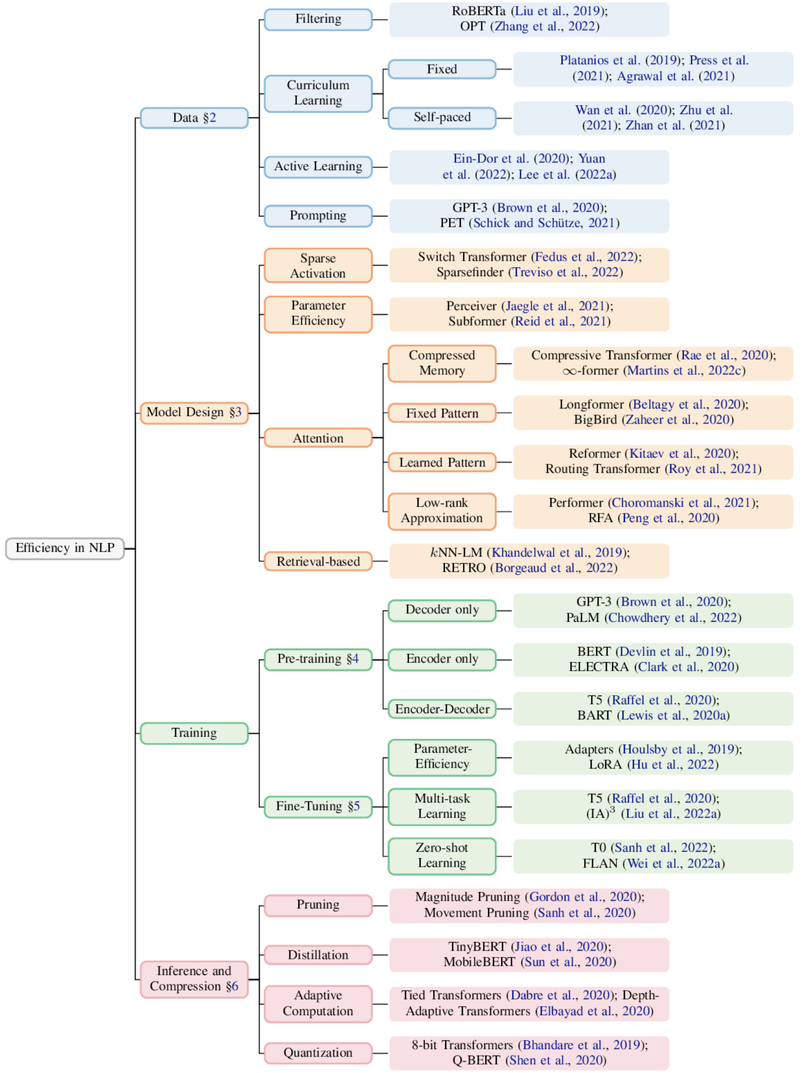


Efficient Modeling

Open Questions

- Can we find the next generation of Transformers?
  - S4 (Gu et al., 2021)
- Should we store knowledge in the model parameters?
  - Retrieval-based models
  - Gu et al (2018); Lewis et al. (2020); Li et al. (2022); Borgeaud et al. (2022)

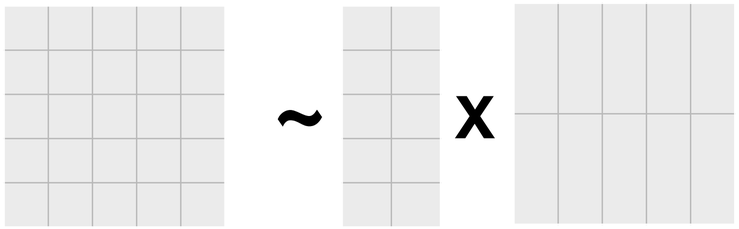
Efficient Methods in NLP



- Efficiency in NLP
  - Training
    - Pre-training §4
      - Decoder only
        - GPT-3 (
        - PaLM (Chowdhery et al., 20
      - Encoder only
        - BERT (Devlin et al., 2019)
        - ELECTRA (Clark et al., 202
      - Encoder-Decoder
        - T5 (Raffel et al., 2020);
        - BART (Lewis et al., 2020a
    - Fine-Tuning §5
      - Parameter-Efficiency
        - Adapters (Houlsby et al., 201
        - LoRA (Hu et al., 2022)
      - Multi-task Learning
        - T5 (Raffel et al., 2020);
        - $(IA)^3$ (Liu et al., 2022a)
      - Zero-shot Learning
        - T0 (Sanh et al., 2022);
        - FLAN (Wei et al., 2022a)
  - Inference and Compression §6
    - Pruning
      - Magnitude Pruning (Gordon et al., 2020);
      - Movement Pruning (Sanh et al., 2020)
    - Distillation
      - TinyBERT (Jiao et al., 2020);
      - MobileBERT (Sun et al., 2020)
    - Adaptive Computation
      - Tied Transformers (Dabre et al., 2020); Depth-Adaptive Transformers (Elbayad et al., 2020)
  - Retrieval-based
    - kNN-LM (Khandelwal et al., 201
    - RETRO (Borgeaud et al., 2022)

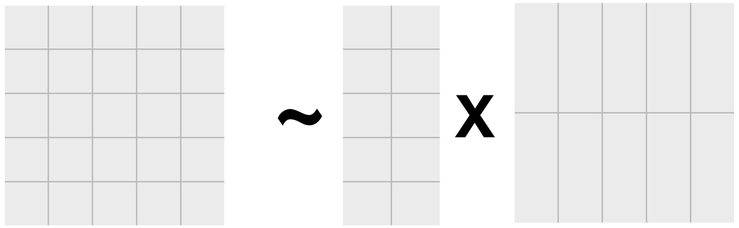
Space Efficiency

- Weight Factorization
  - Lan et al. (2019); Wang et al. (2019)

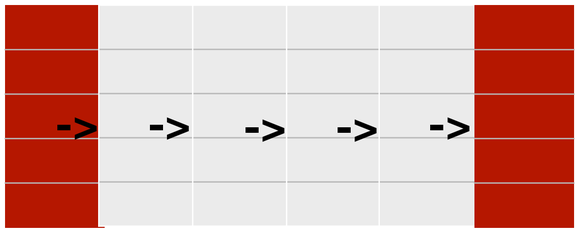


Space Efficiency

- Weight Factorization
  - Lan et al. (2019); Wang et al. (2019)



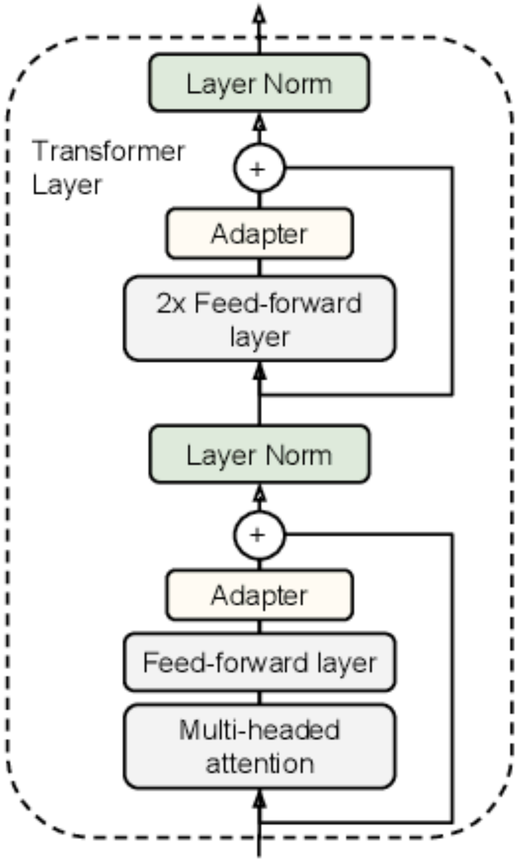
- Weight Sharing
  - Inan et al. (2016); Press & Wolf (2017); Dehghani et al. (2019)



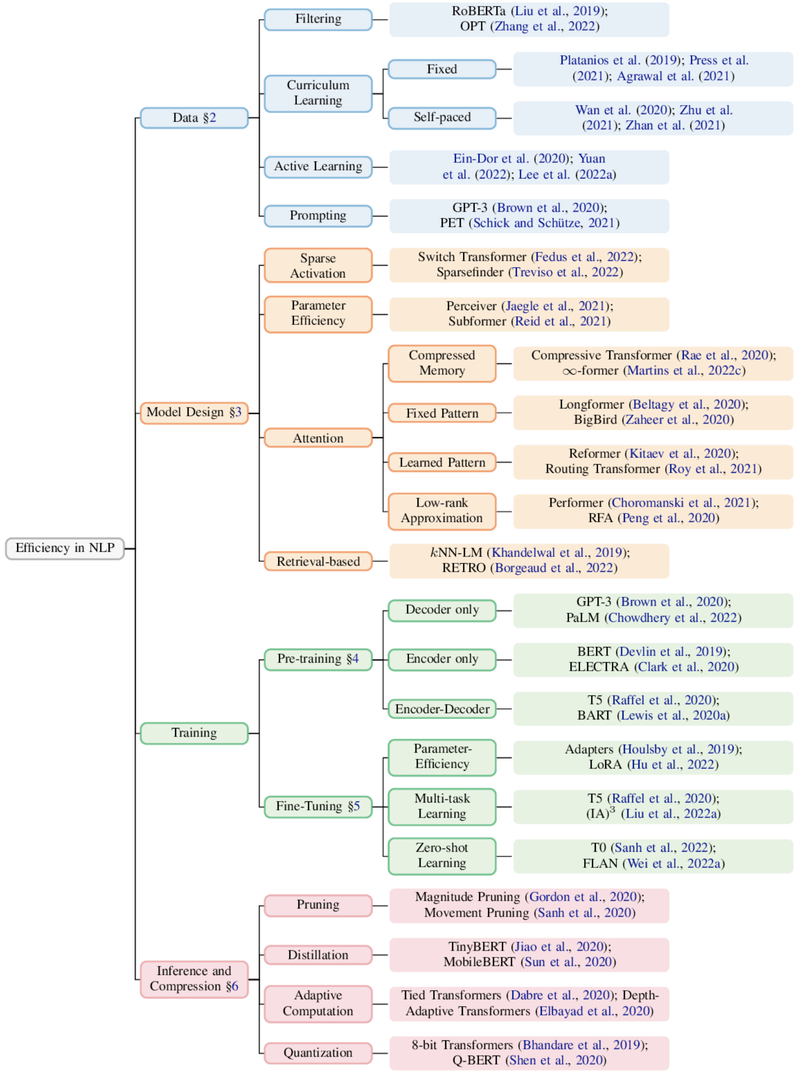
Adapters

Houlsby et al. (2019)

- Instead of updating all parameters during training, add small components (**adapters**) and only update those



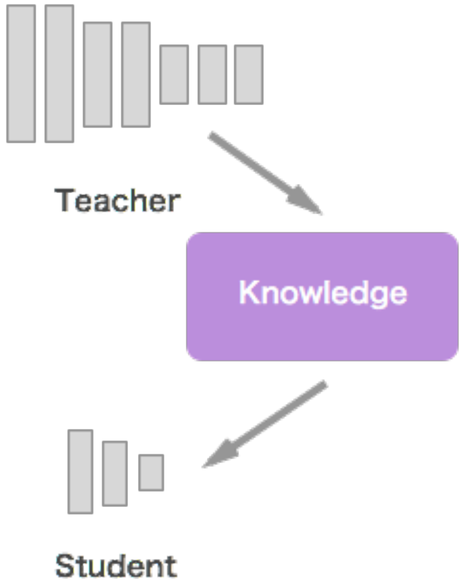
Efficient Methods in NLP



- Efficiency
- Fine-Tuning §5
  - Multi-task Learning
    - T5 (
    - (IA)
  - Zero-shot Learning
    - T0 (Sanh et al., 2022);
    - FLAN (Wei et al., 2022a
- Inference and Compression §6
  - Pruning
    - Magnitude Pruning (Gordon et al., 2020);
    - Movement Pruning (Sanh et al., 2020)
  - Distillation
    - TinyBERT (Jiao et al., 2020);
    - MobileBERT (Sun et al., 2020)
  - Adaptive Computation
    - Tied Transformers (Dabre et al., 2020); Depth-Adaptive Transformers (Elbayad et al., 2020)
  - Quantization
    - 8-bit Transformers (Bhandare et al., 2019);
    - Q-BERT (Shen et al., 2020)

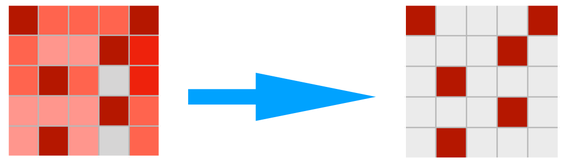
Model Distillation

- Student-teacher model
  - Sun et al. (2019); Sanh et al. (2019); Zhao et al (2019)
- Student performs on par with teacher
  - Better than a student of similar size trained from scratch



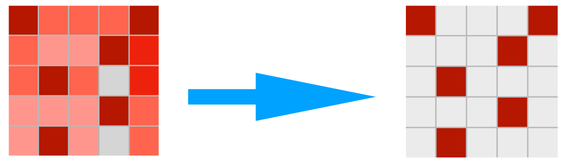
Pruning

- Learn a sparse version of a trained model
  - Gale et al. (2019); Lee et al. (2019); Frankle & Corbin (2019)



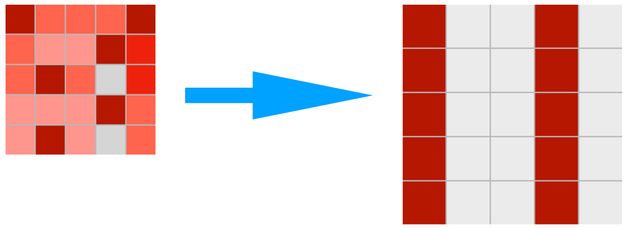
Pruning

- Learn a sparse version of a trained model
  - Gale et al. (2019); Lee et al. (2019); Frankle & Corbin (2019)
- Hard to translate to speed and energy savings
  - Although see BlockSparse (Gray et al., 2017)
  - Potentially useful for fast training (Dettmers & Zettlemoyer, 2019)



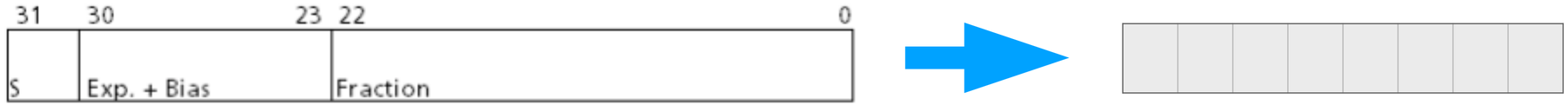
Structured Pruning

- Learn a sparse structure of the trained model
  - I.e., prune attention heads, layers, etc.
  - Gordon et al. (2018); Dodge, **S.** et al. (2019); Michel et al. (2019); Wang et al. (2019); Fan et al. (2020)
- Could be leveraged to **perform fewer computations**



Quantization

- Reduced precision
  - Zafrir et al. (2019); Shen et al. (2019); Bhandare et al. (2019)

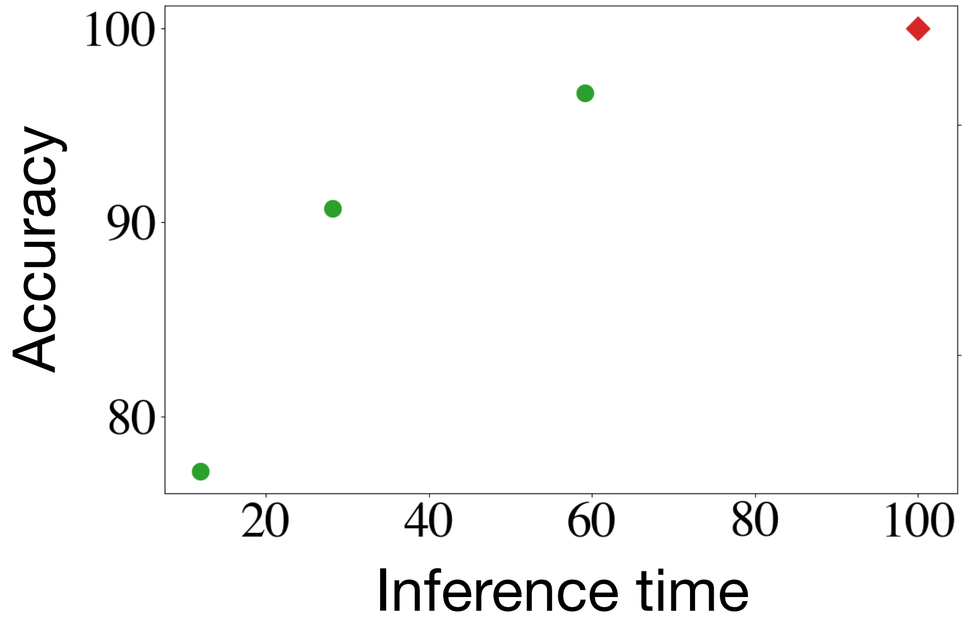


Matching Model and Instance Complexity

S. et al., ACL 2020

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

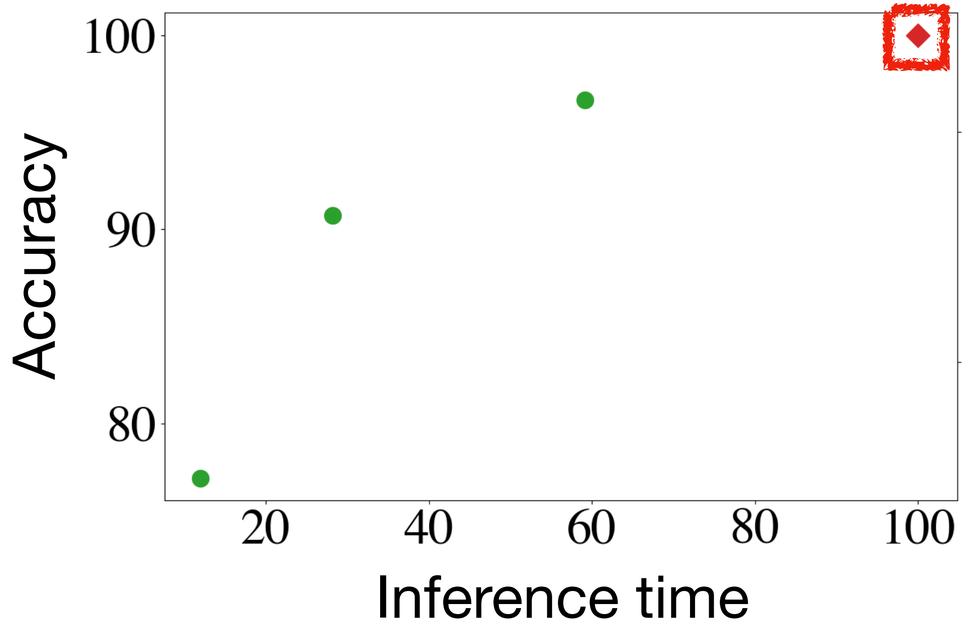


Matching Model and Instance Complexity

S. et al., ACL 2020

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

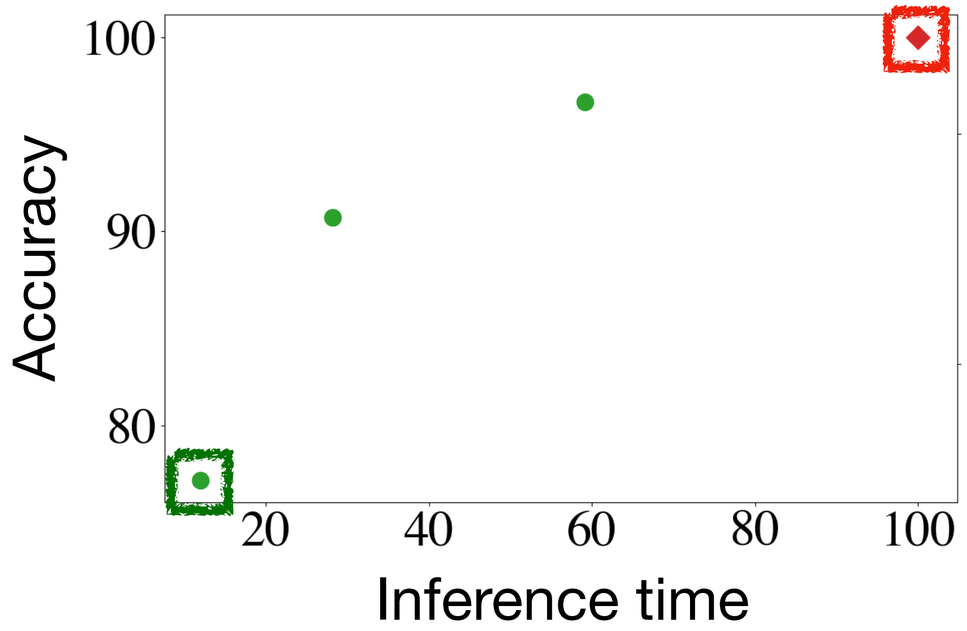


Matching Model and Instance Complexity

S. et al., ACL 2020

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

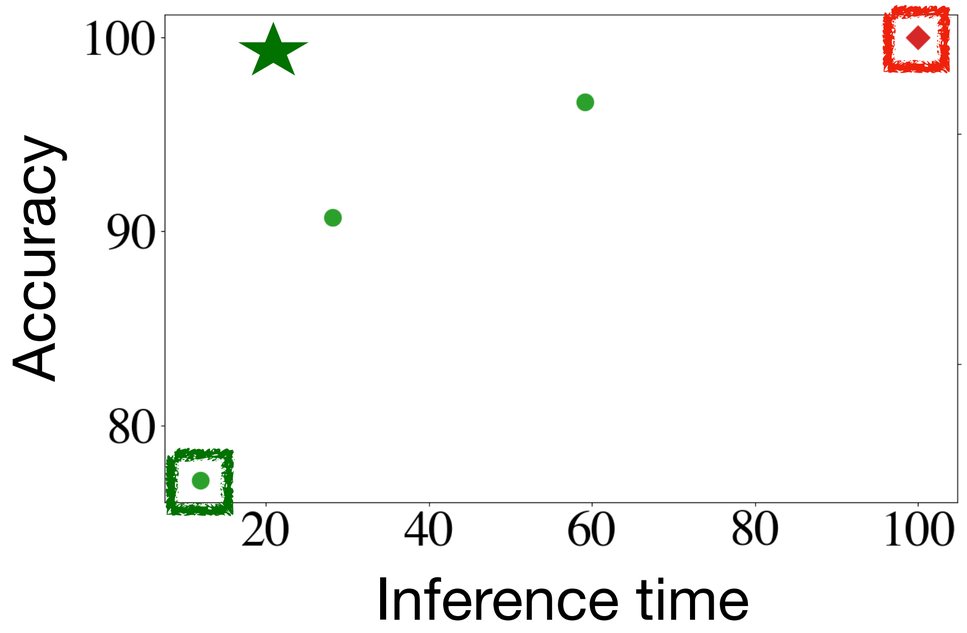


Matching Model and Instance Complexity

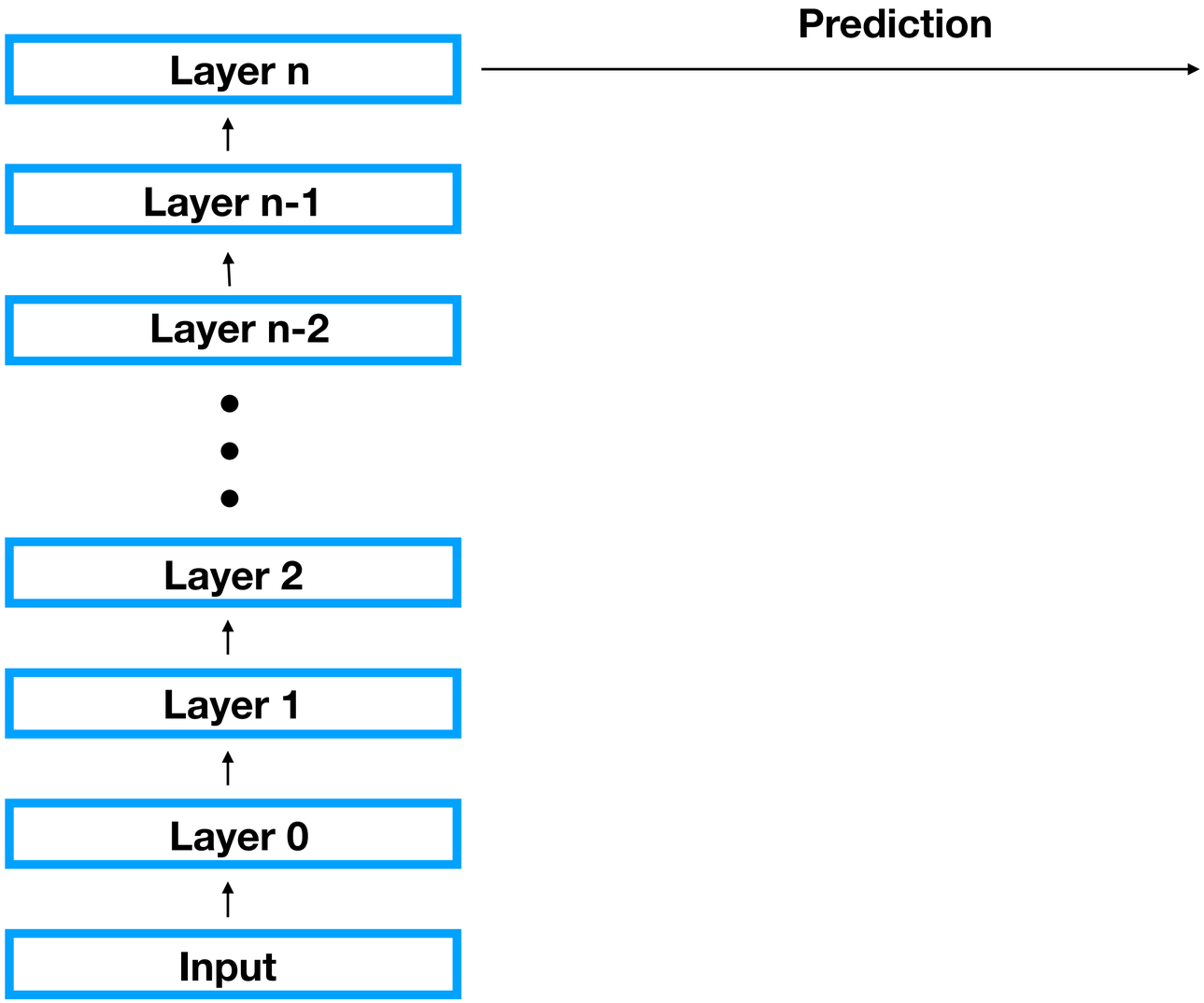
S. et al., ACL 2020

Run an efficient model on "easy" instances,

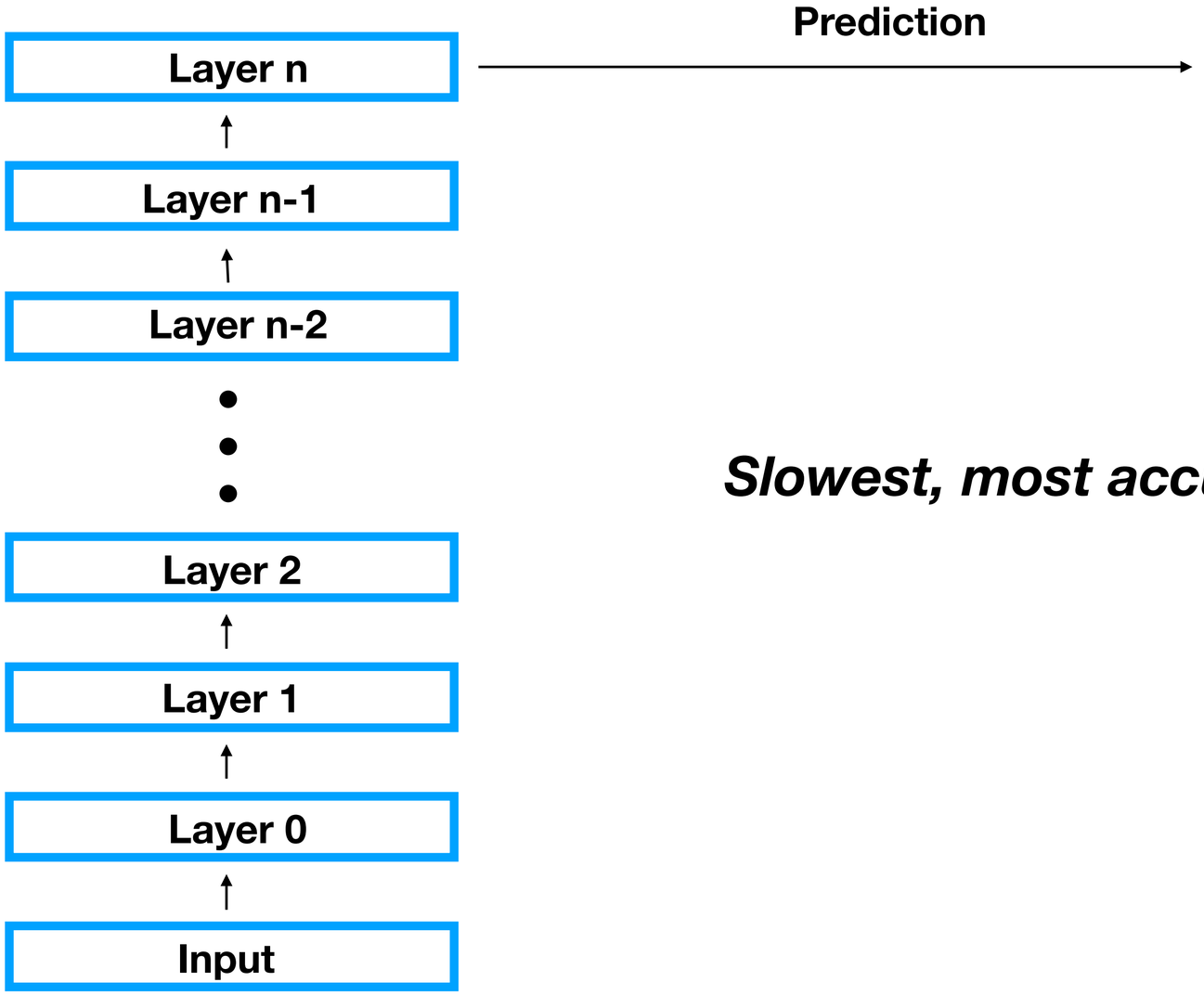
and an expensive model on "hard" instances



Pretrained BERT Fine-tuning

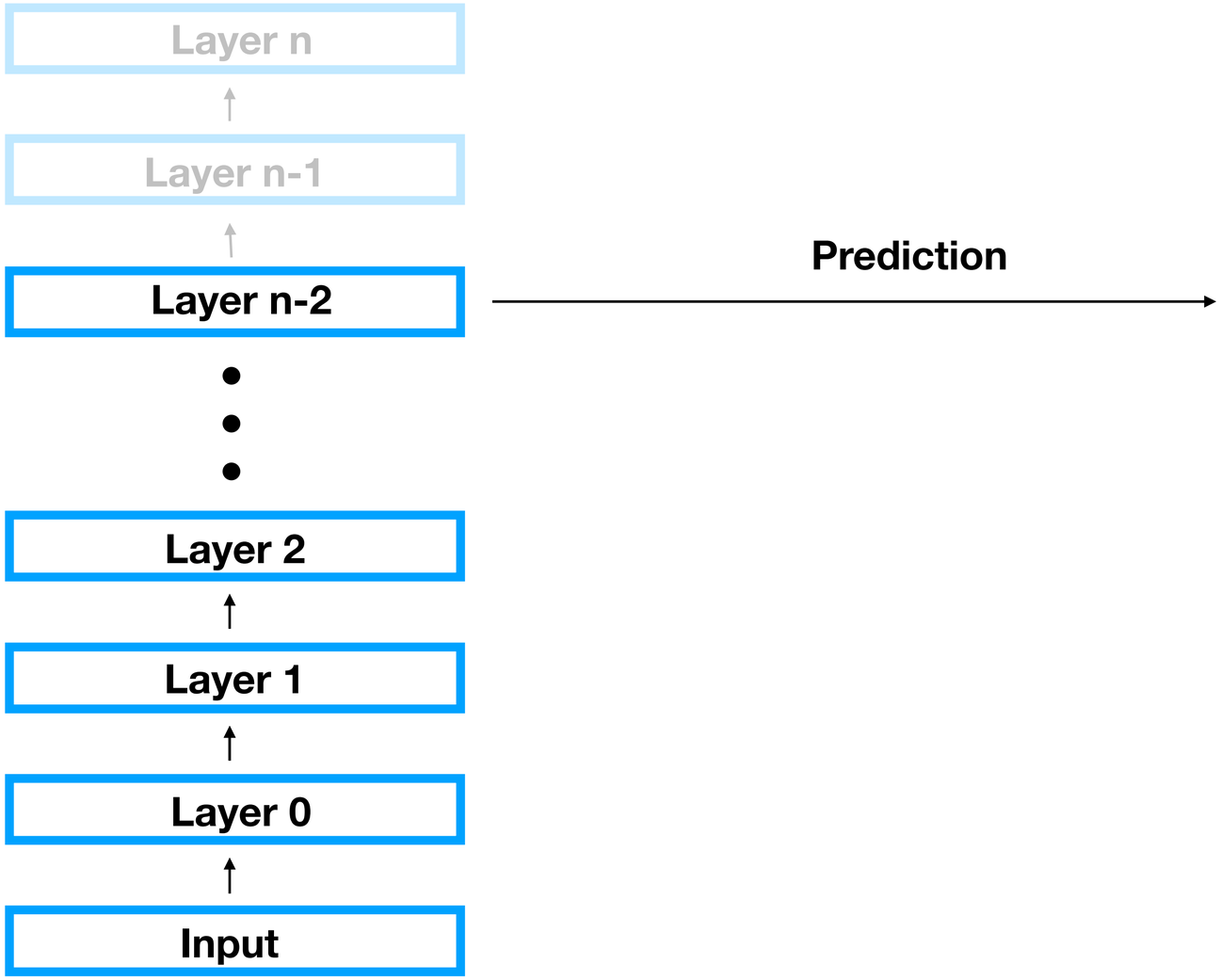


Pretrained BERT Fine-tuning

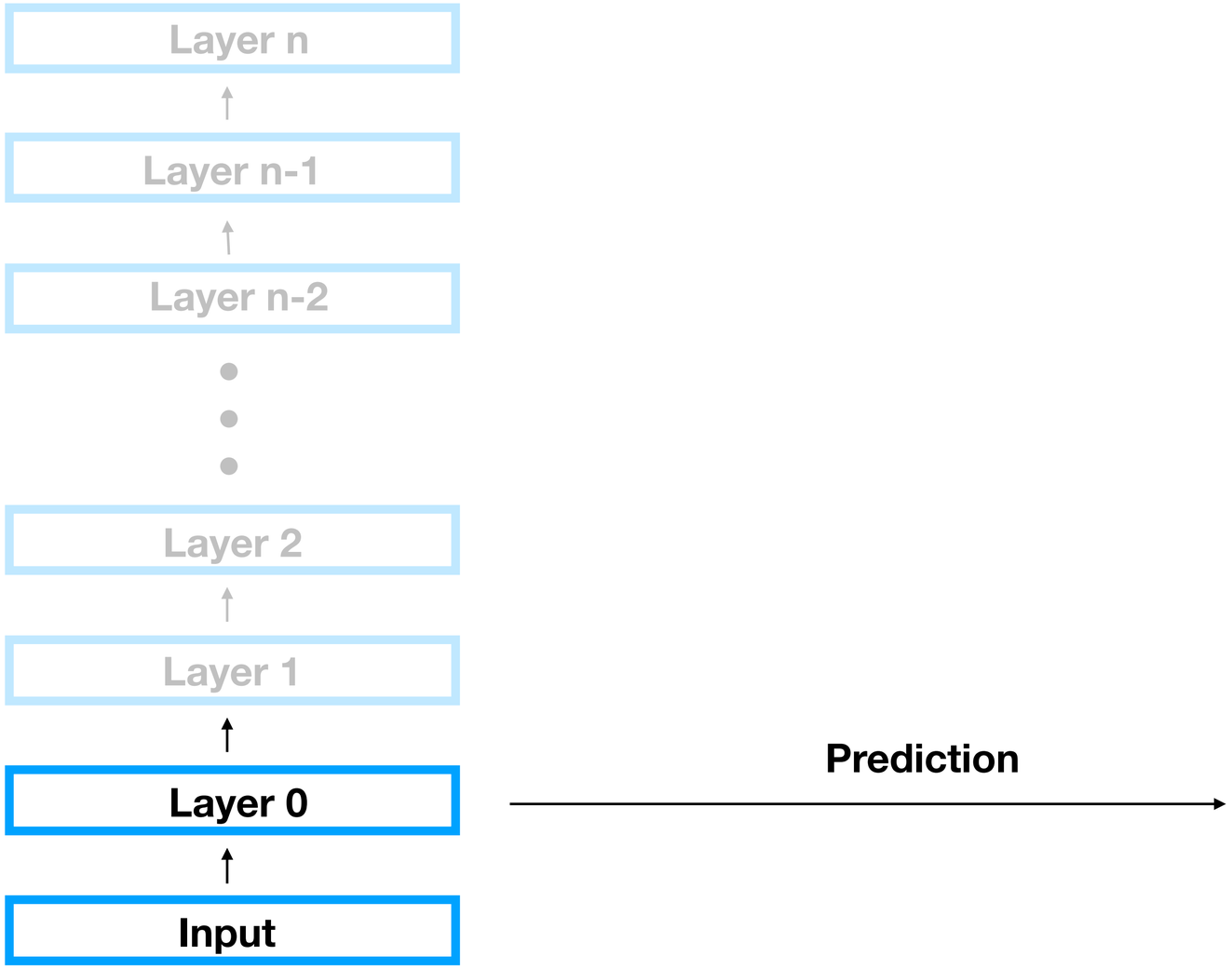


Slowest, most accurate

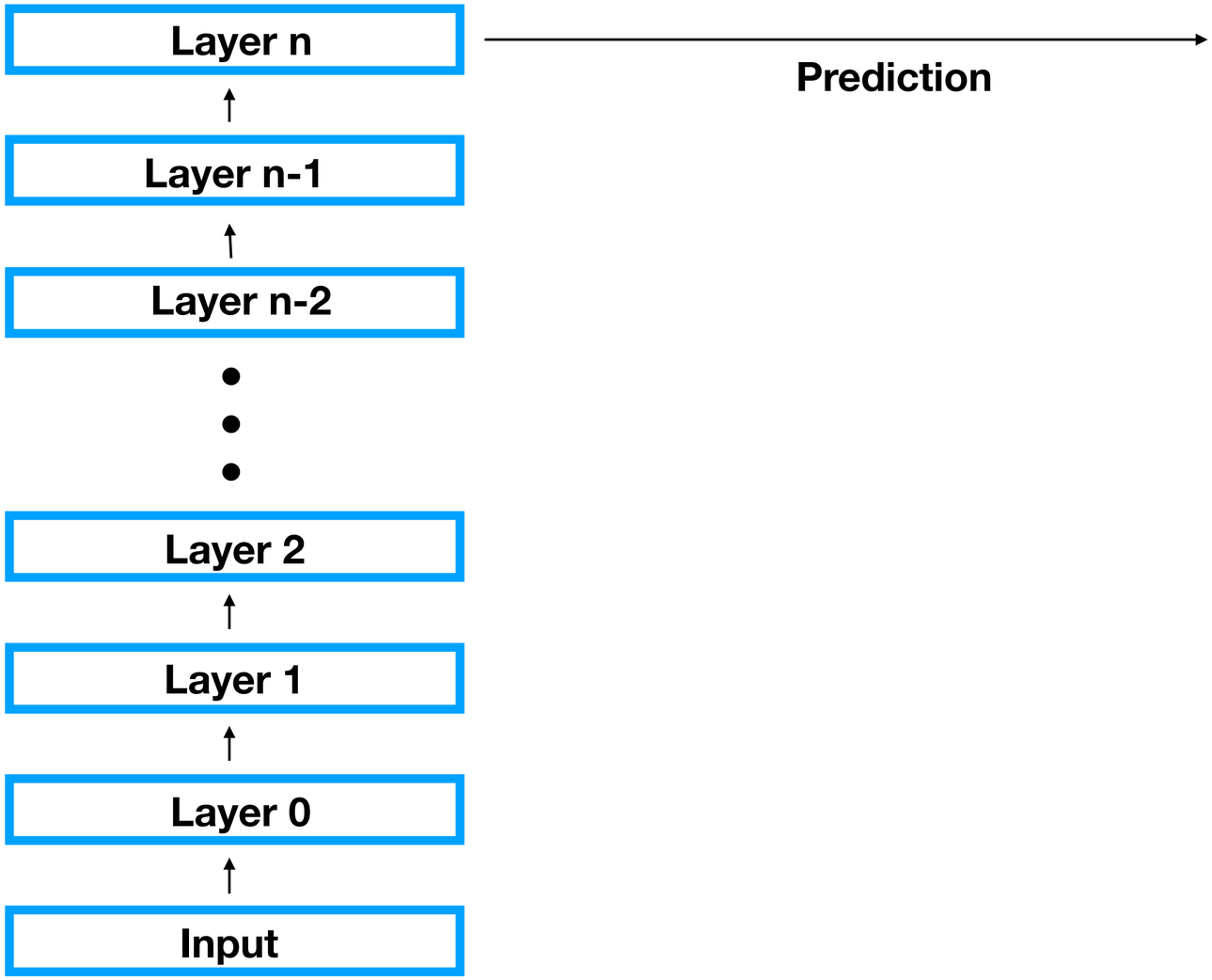
Faster, less Accurate



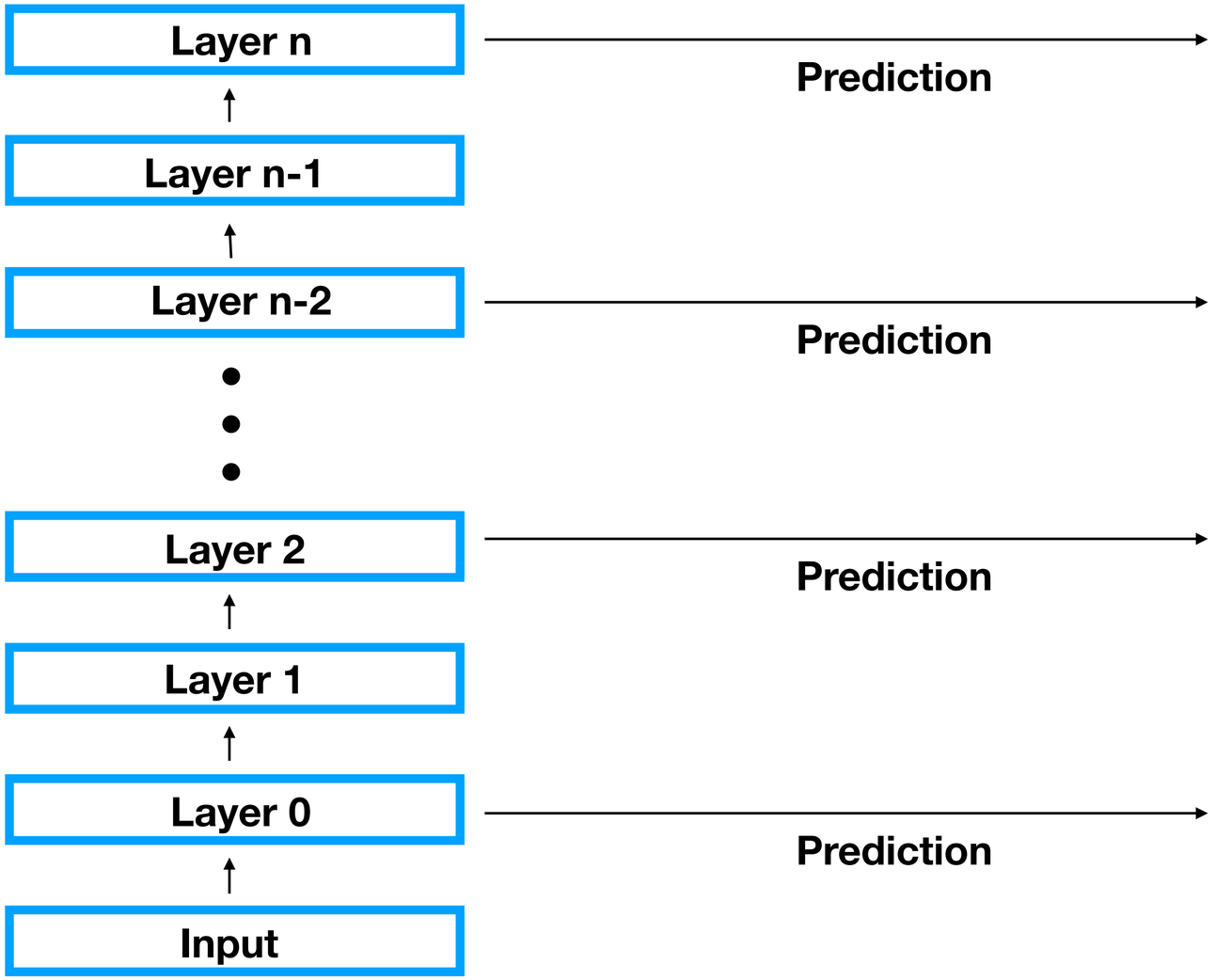
Fastest, least Accurate



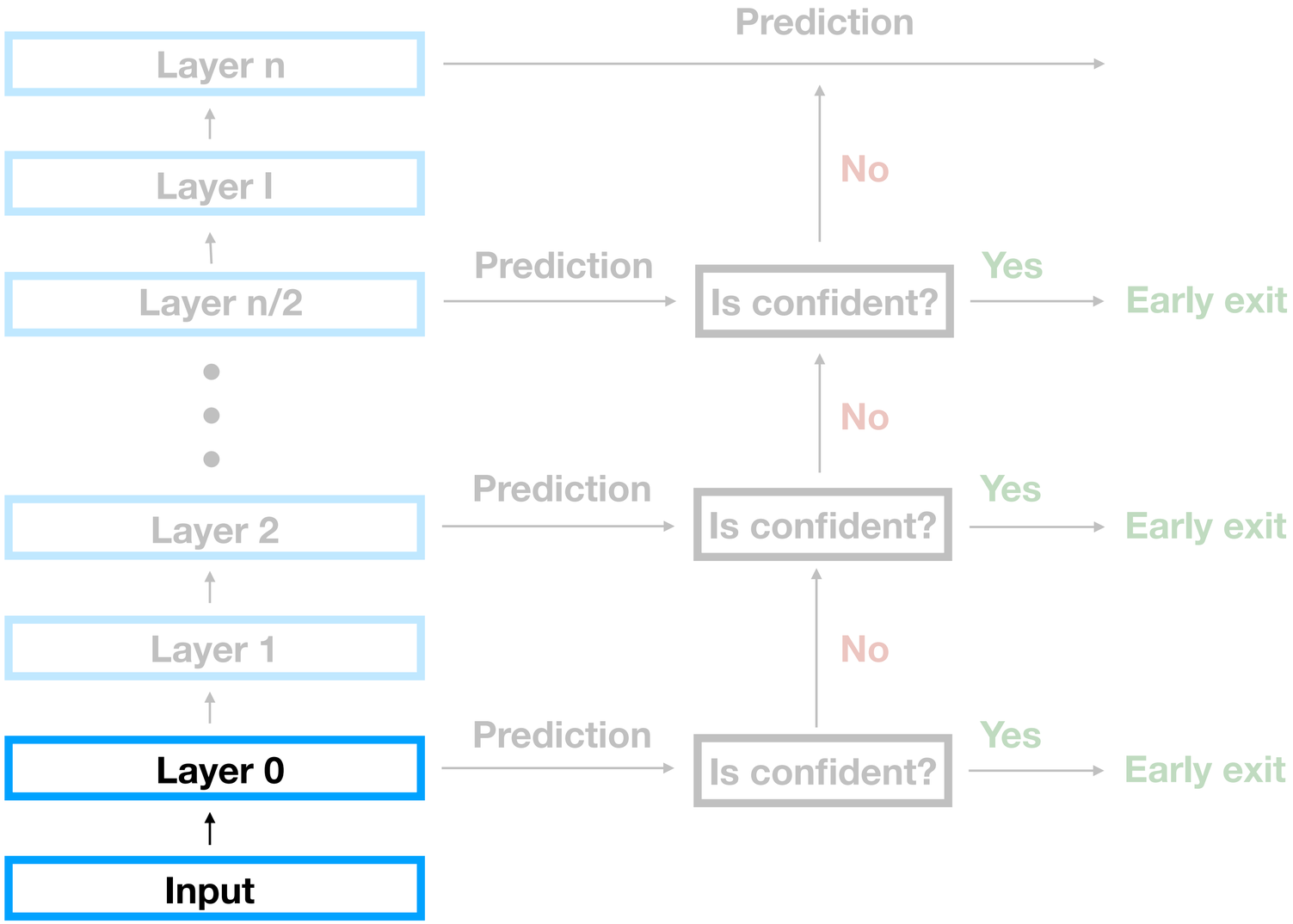
Our Approach: **Training** Time



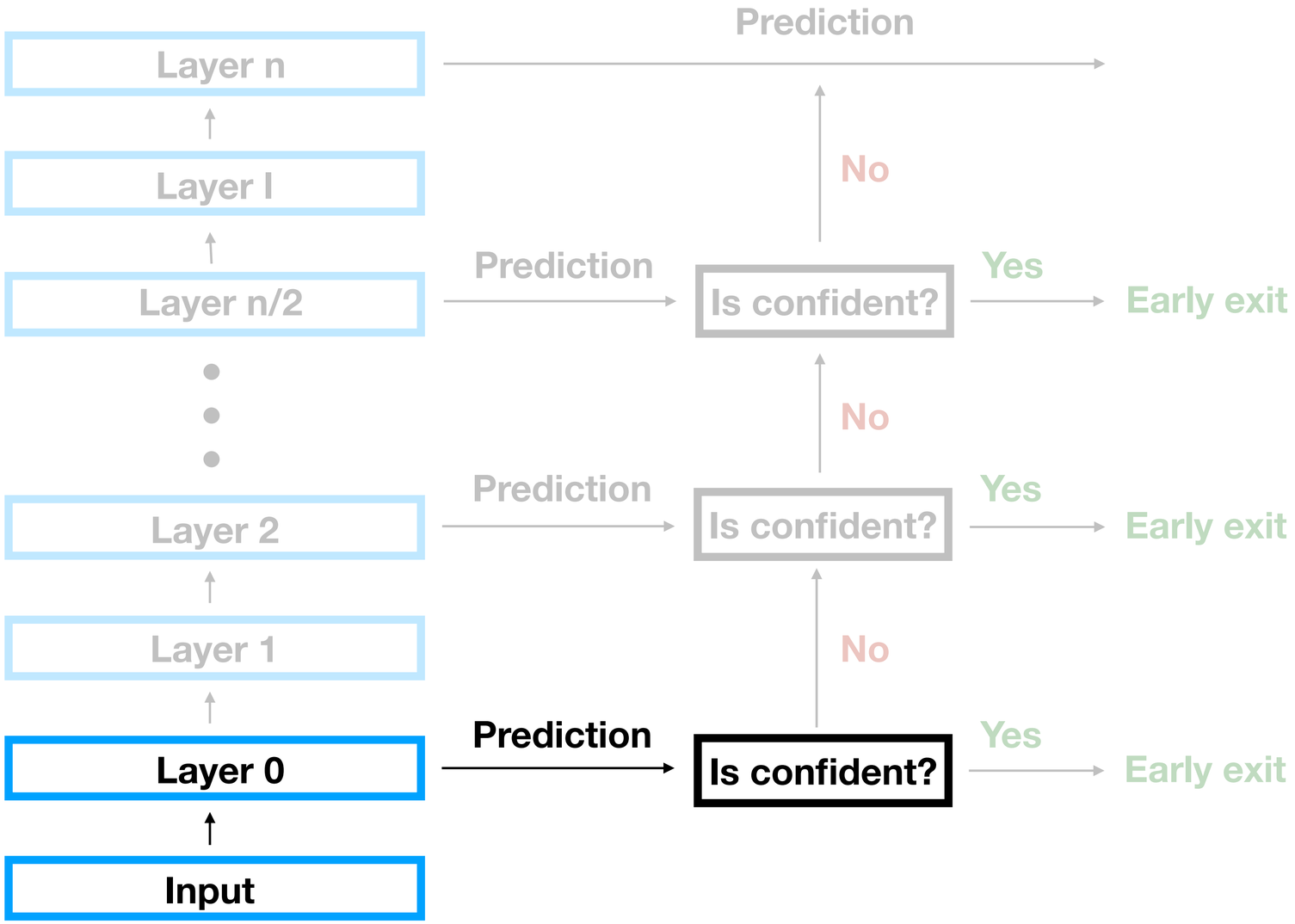
Our Approach: **Training** Time



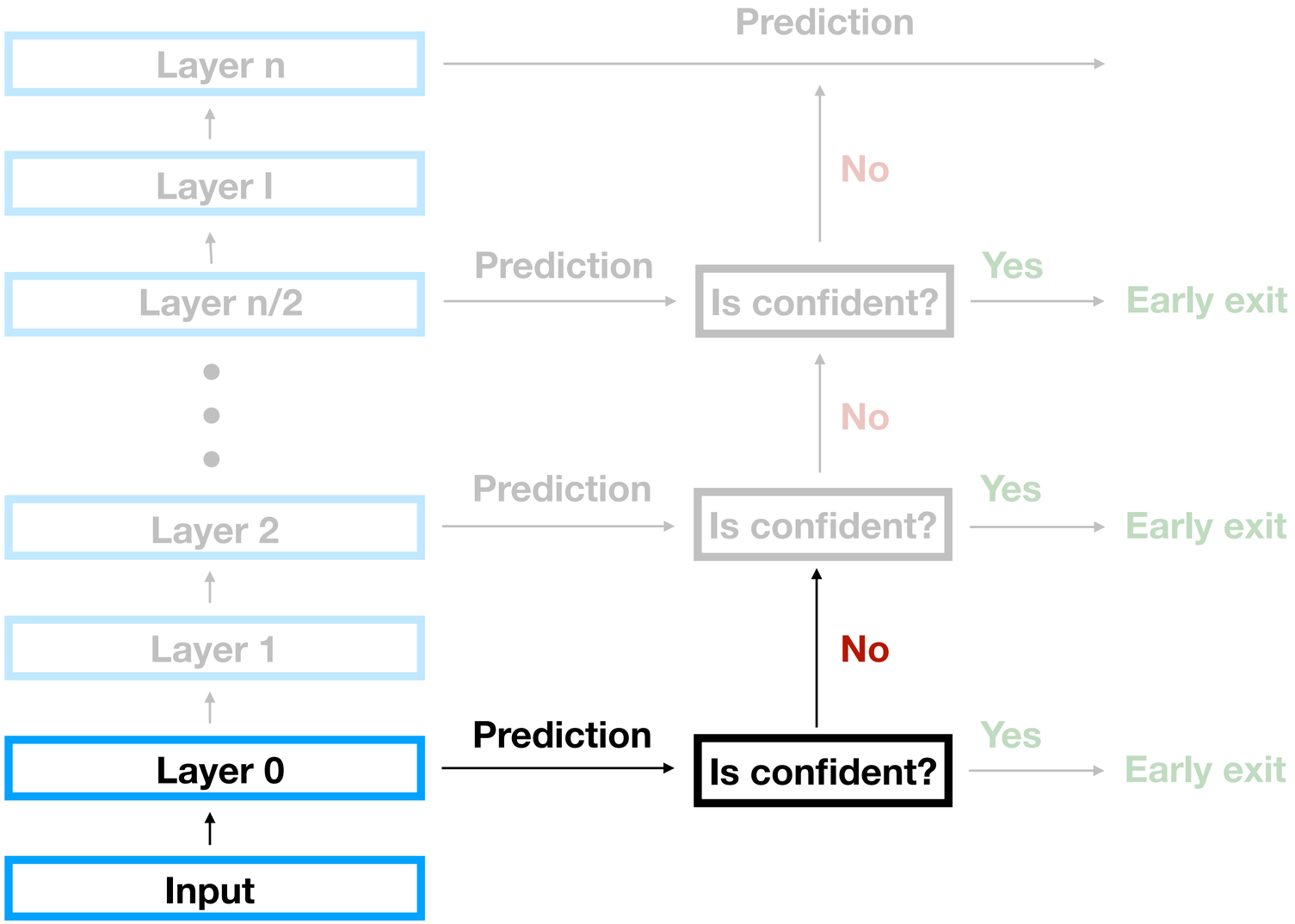
Our Approach: Test Time



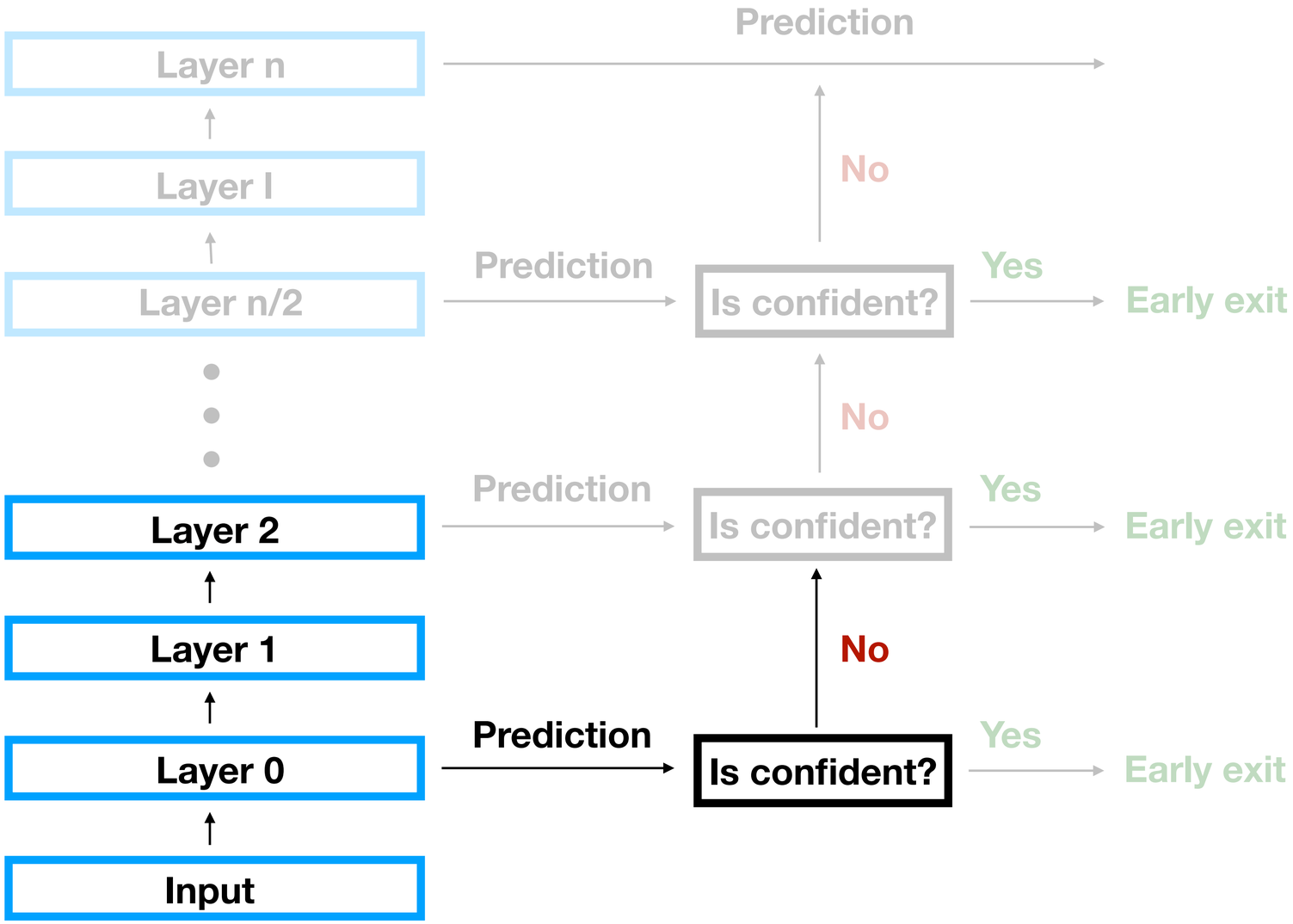
Our Approach: Test Time



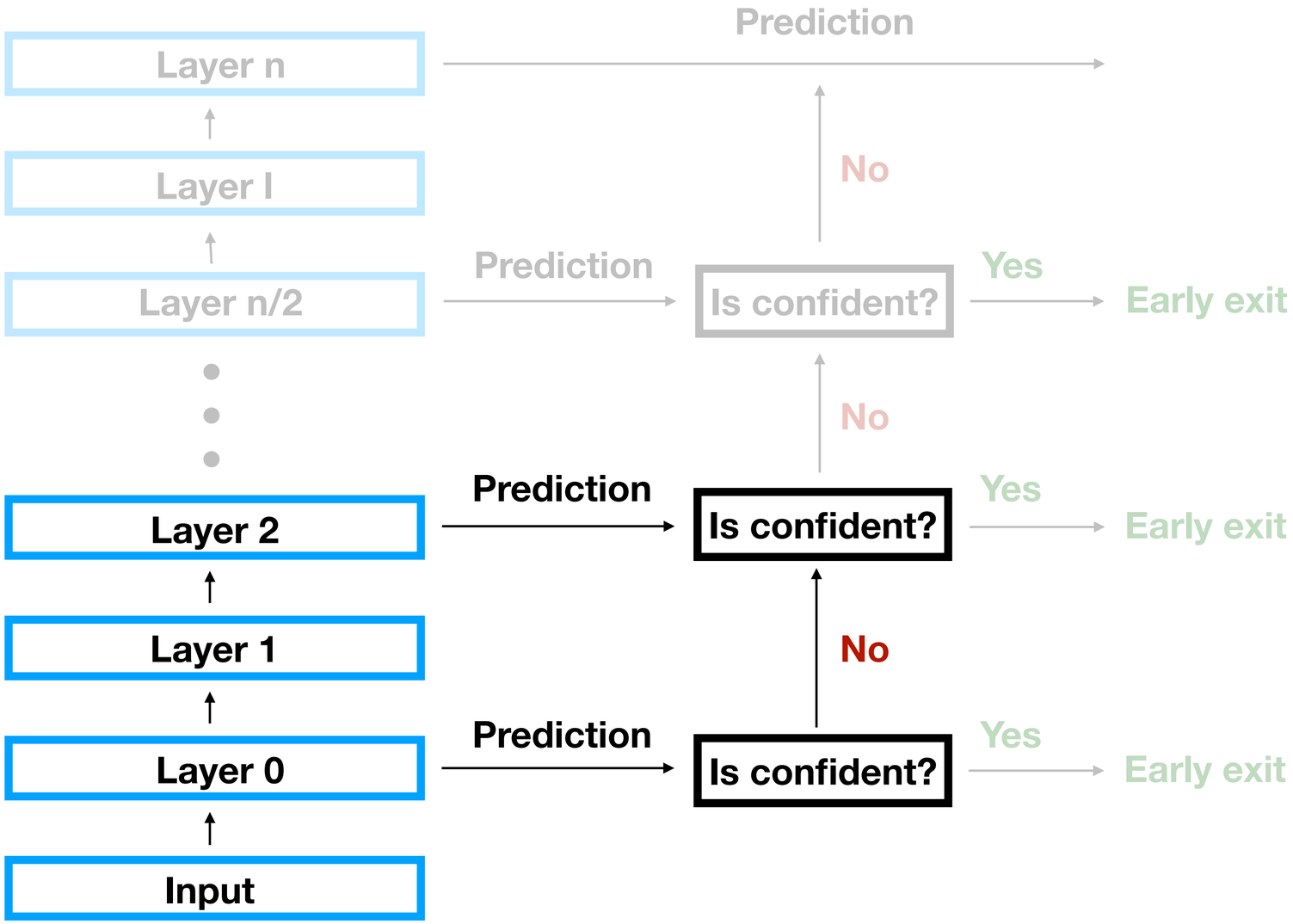
Our Approach: Test Time



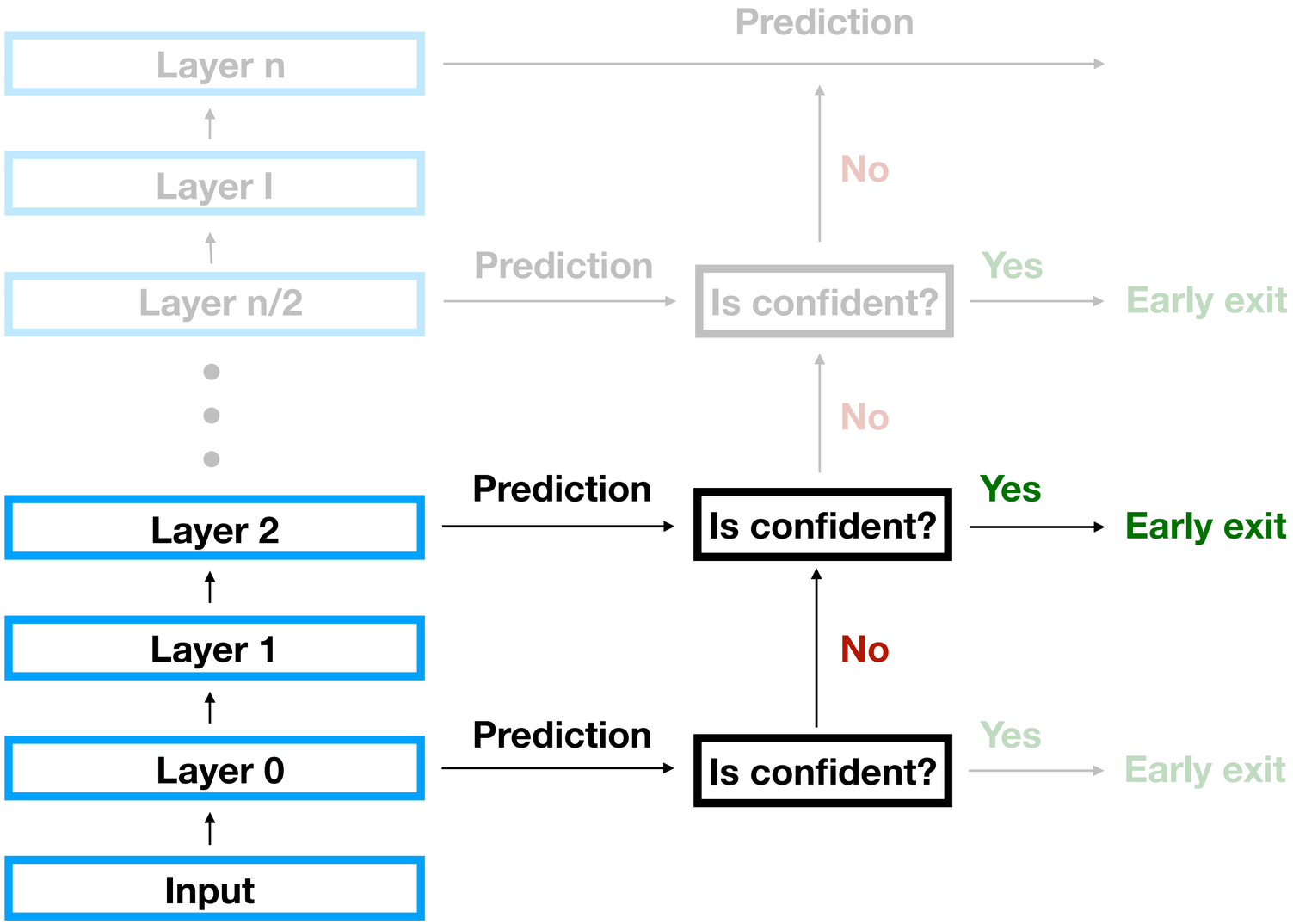
Our Approach: Test Time



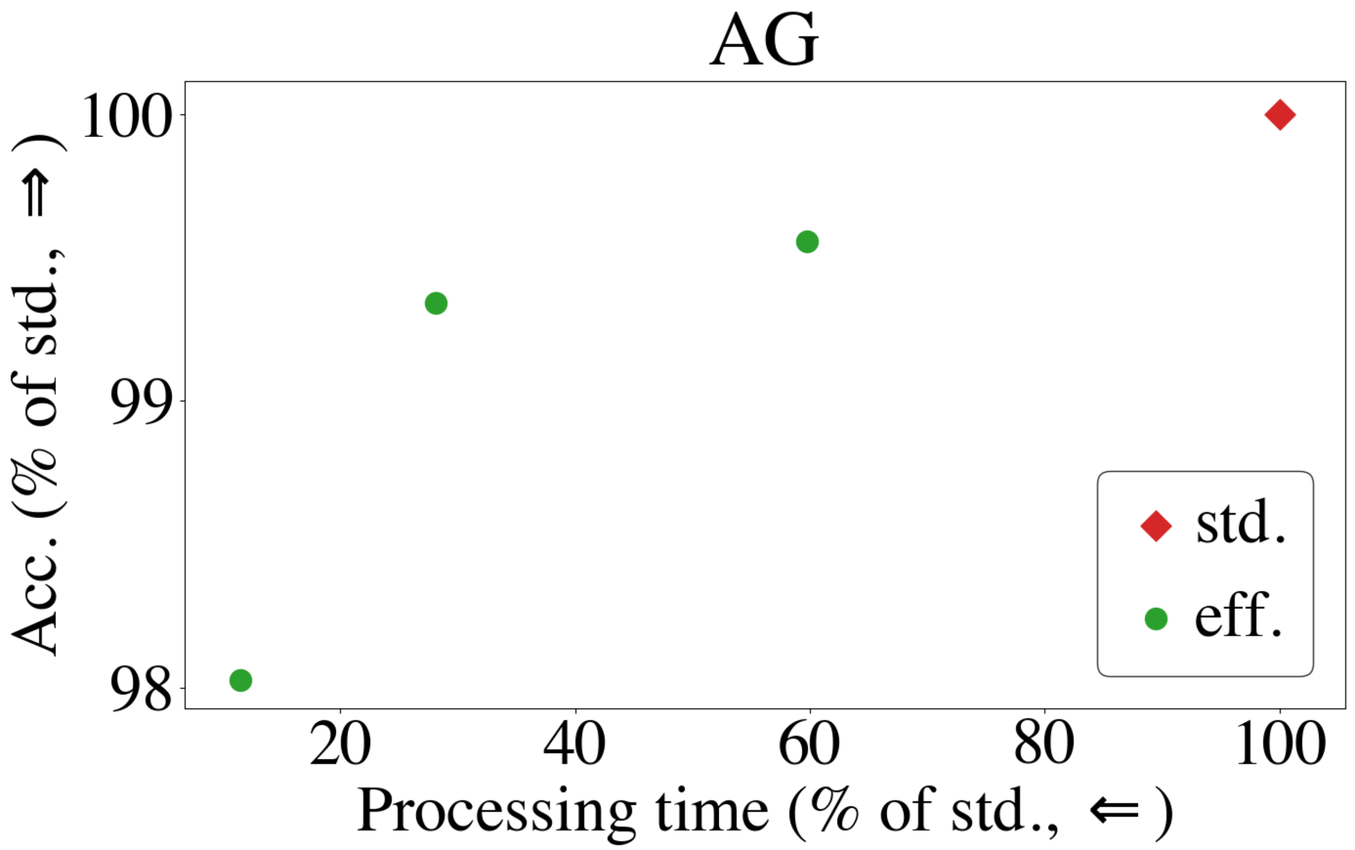
Our Approach: Test Time



Our Approach: Test Time

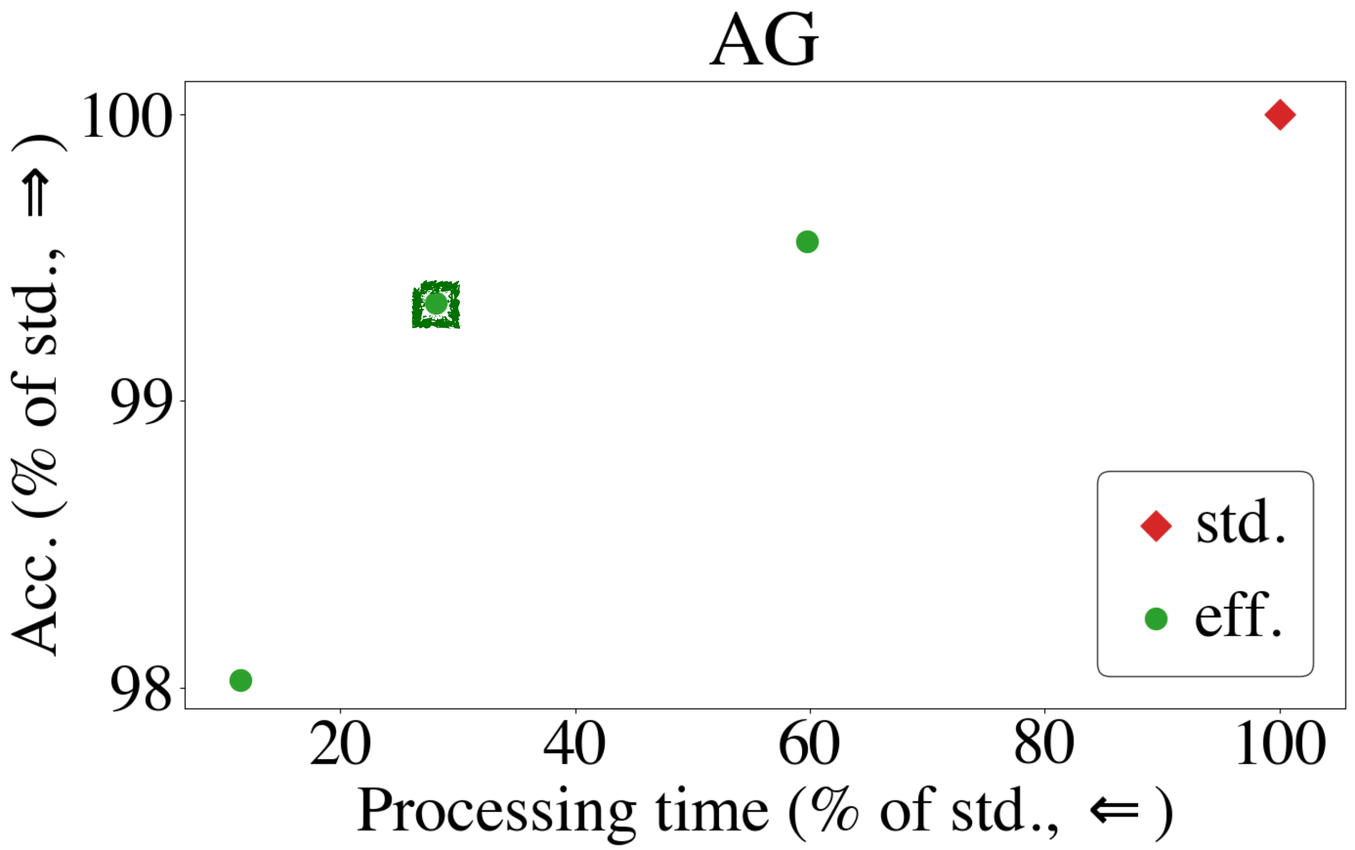


Strong Baselines!



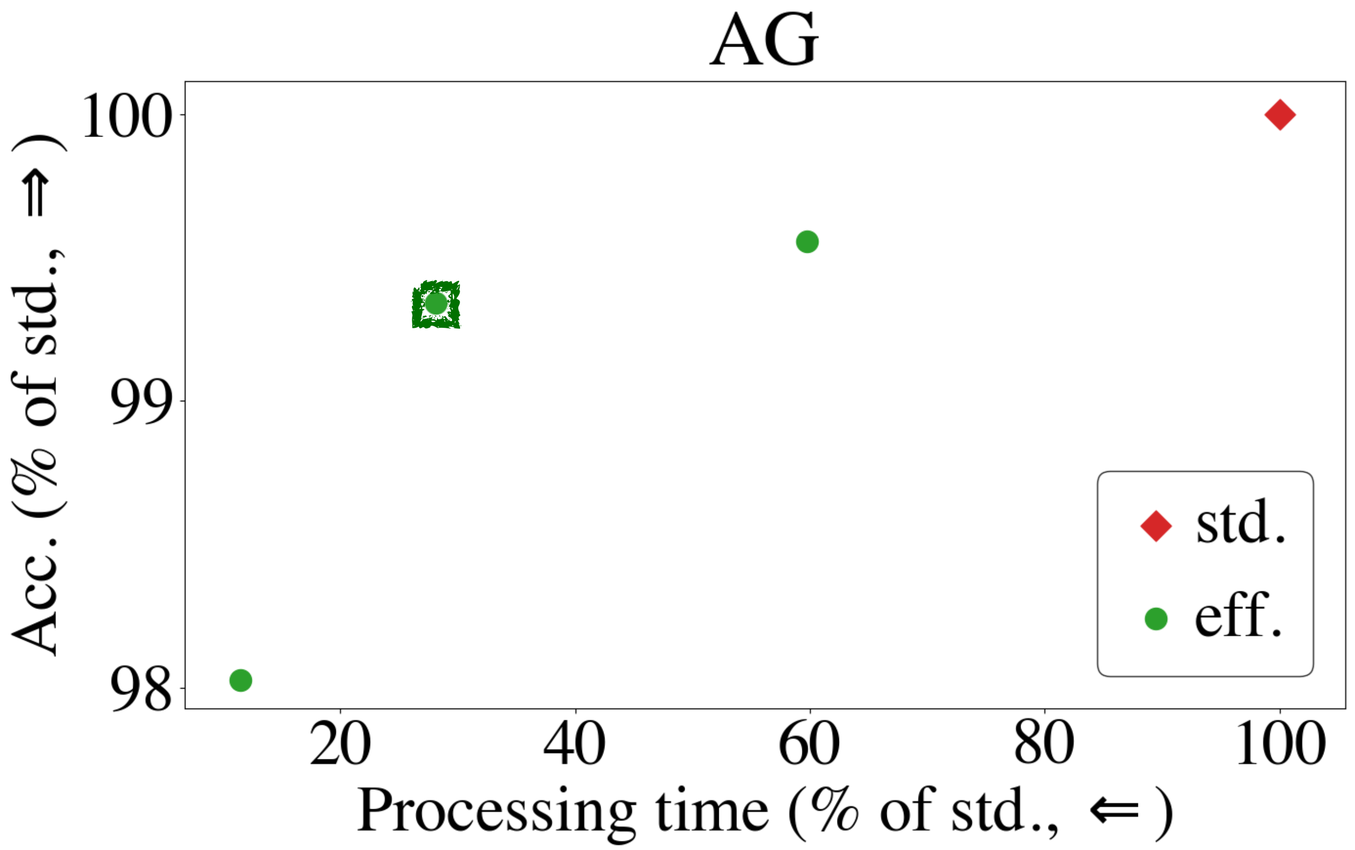
69

Strong Baselines!



69

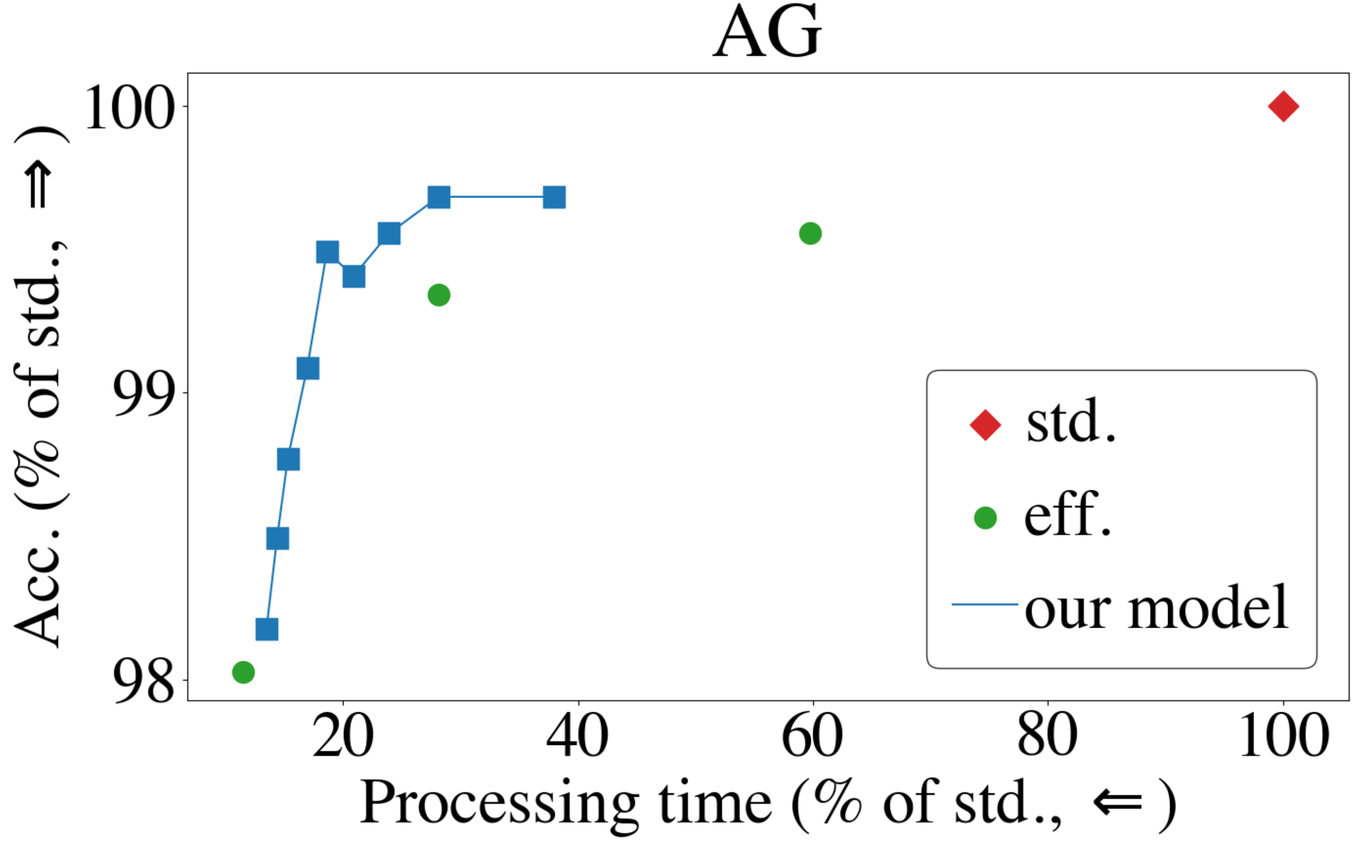
Strong Baselines!



3 times faster, within 1% of full model

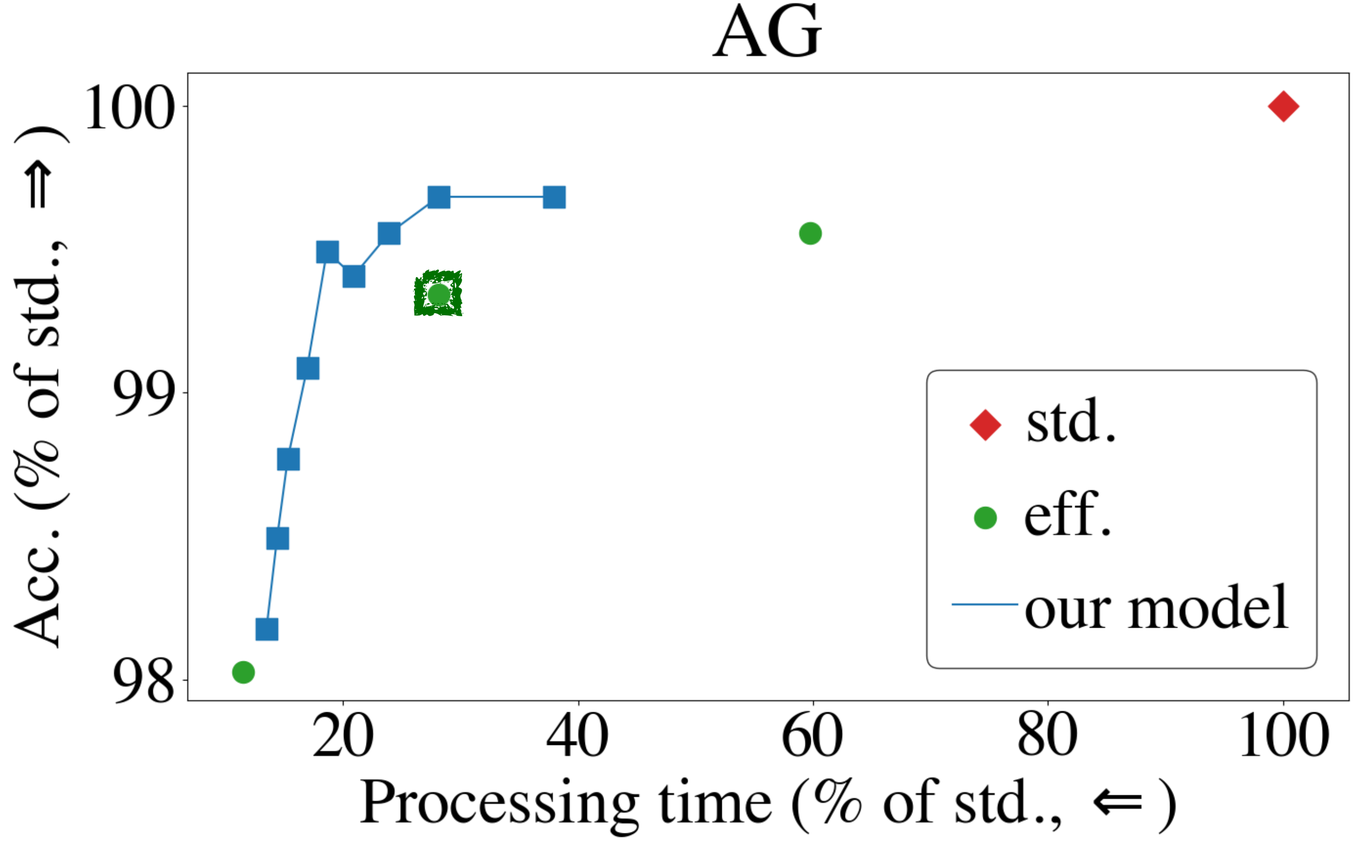
69

Better Speed/Accuracy Tradeoff



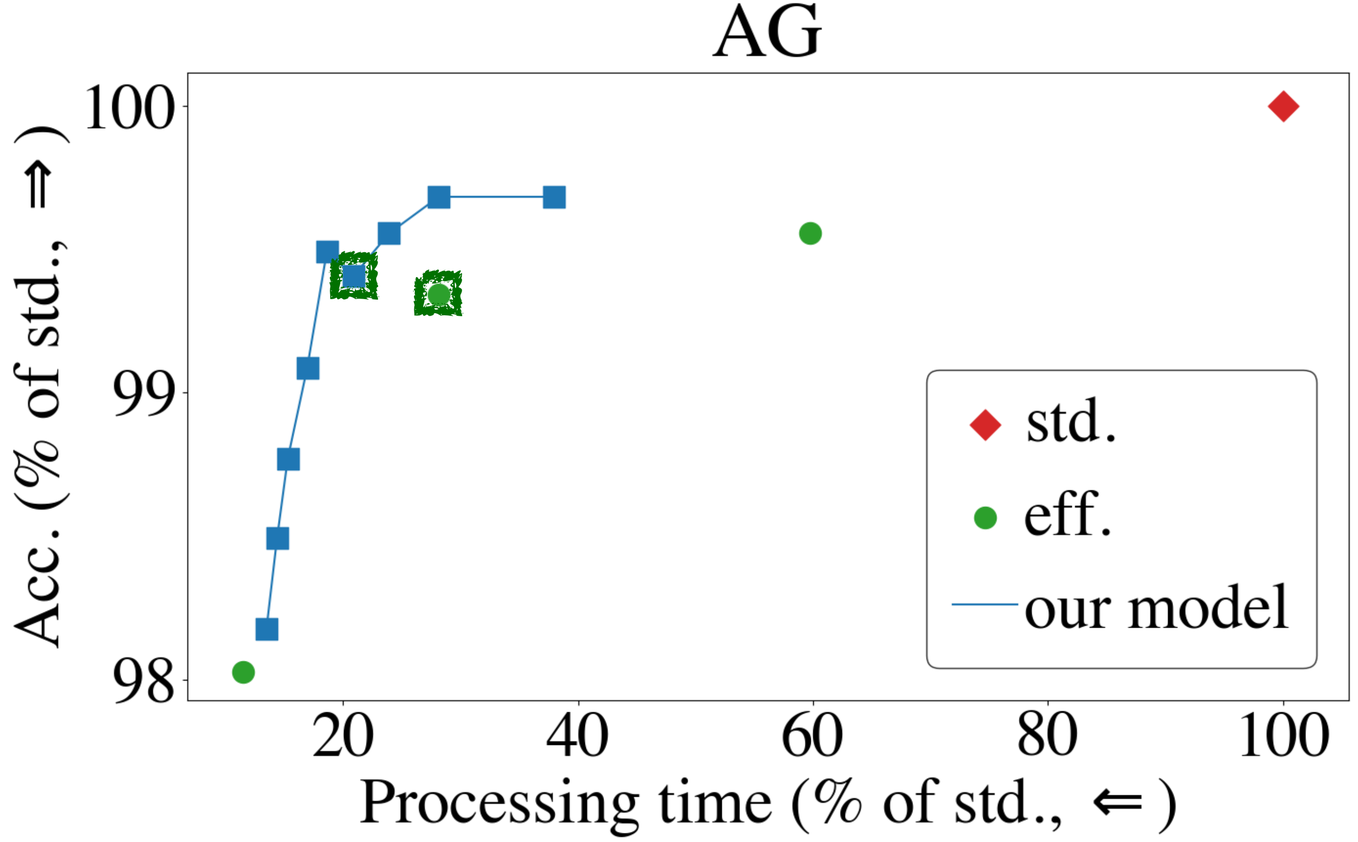
70

Better Speed/Accuracy Tradeoff



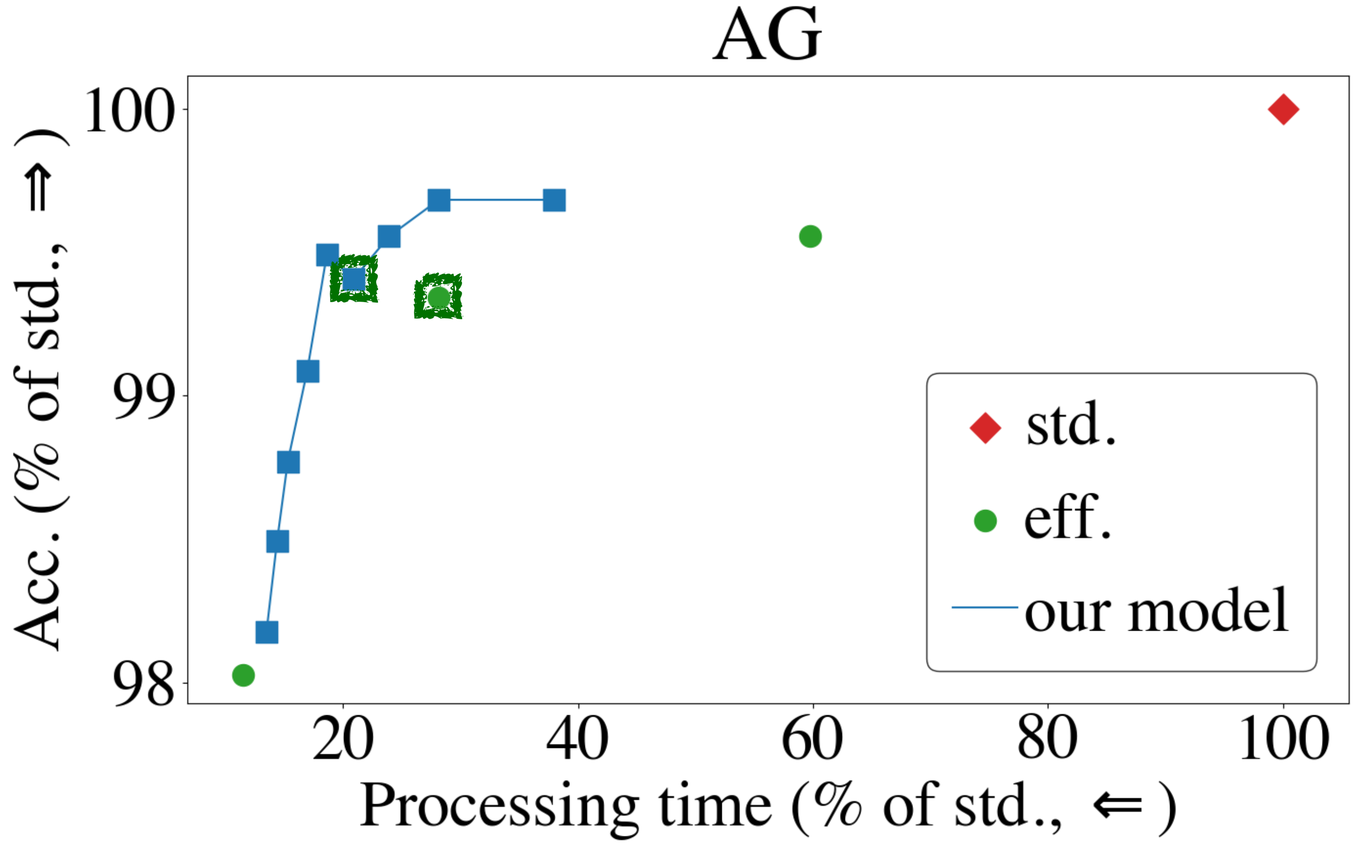
70

Better Speed/Accuracy Tradeoff



70

Better Speed/Accuracy Tradeoff



5 times faster, within 1% of full model

70

Efficiency
Open Questions

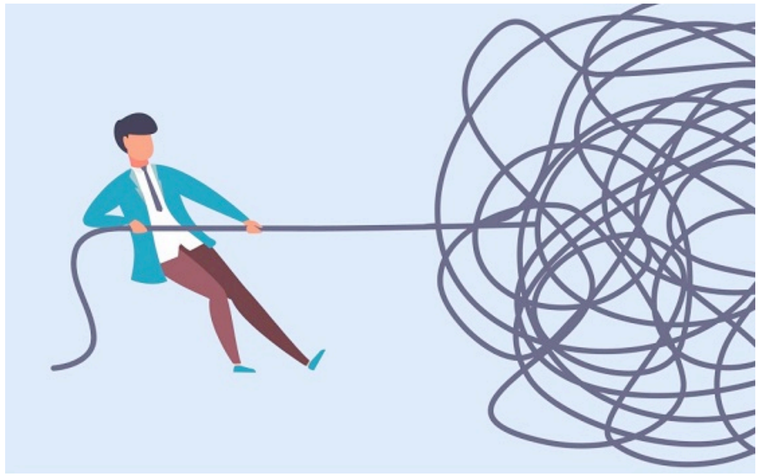
- What makes a good sparse structure?

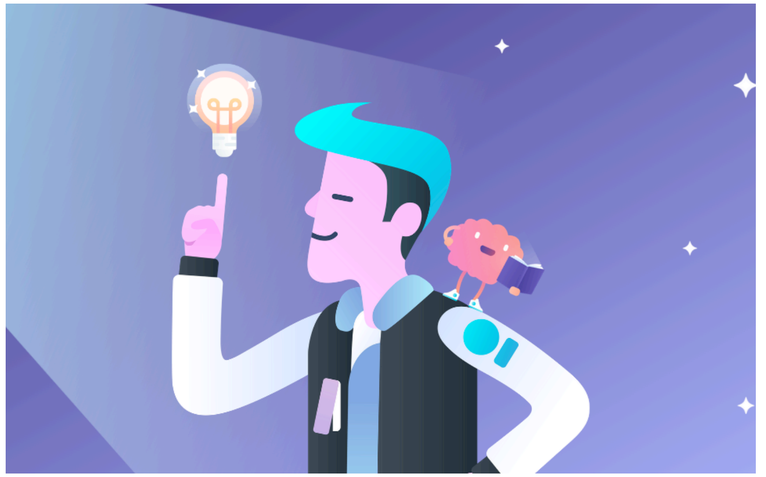
Efficiency

Open Questions

- What makes a good sparse structure?
- Combining different methods

Outline



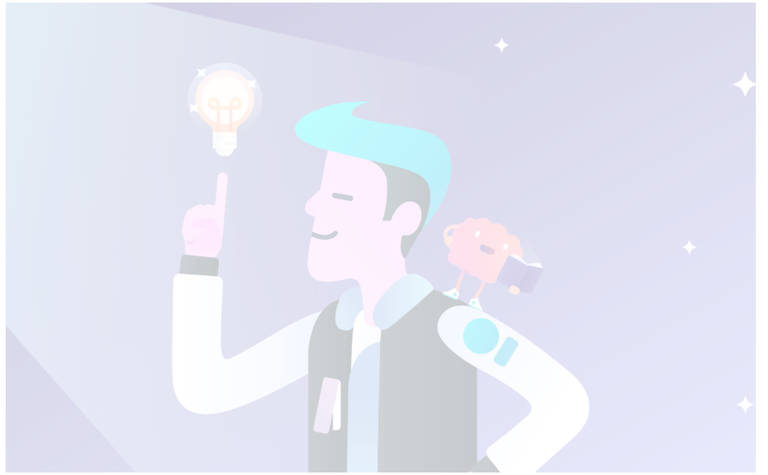


Challenges **Opportunities** **Mitigation**

Outline

Challenges

Opportunities



Mitigation

Tackling Climate Change with Machine Learning

David Rolnick[1]*, Priya L. Donti[2], Lynn H. Kaack[3], Kelly Kochanski[4], Alexandre Lacoste[5], Kris Sankaran[6,7], Andrew Slavin Ross[9], Nikola Milojevic-Dupont[10,11], Natasha Jaques[12], Anna Waldman-Brown[12], Alexandra Luccioni[6,7], Tegan Maharaj[6,8], Evan D. Sherwin[2], S. Karthik Mukkavilli[6,7], Konrad P. Körding[1], Carla Gomes[13], Andrew Y. Ng[14], Demis Hassabis[15], John C. Platt[16], Felix Creutzig[10,11], Jennifer Chayes[17], Yoshua Bengio[6,7]

[1]University of Pennsylvania, [2]Carnegie Mellon University, [3]ETH Zürich, [4]University of Colorado Boulder, [5]Element AI, [6]Mila, [7]Université de Montréal, [8]École Polytechnique de Montréal, [9]Harvard University, [10]Mercator Research Institute on Global Commons and Climate Change, [11]Technische Universität Berlin, [12]Massachusetts Institute of Technology, [13]Cornell University, [14]Stanford University, [15]DeepMind, [16]Google AI, [17]Microsoft Research

	Computer vision	NLP	Time-series analysis	Unsupervised learning	RL & Control	Causal inference	Uncertainty quantification	Transfer learning	Interpretable ML	Other
Electricity Systems	1	1.1	1.1 1.2	1	1.1		1.1 1.2	1.3	1.1	1.1
Transportation	2.1 2.2 2.4		2	2.1 2.4	2	2.1 2.4	2	2.1 2.4	2	
Buildings & Cities	3.2	3.3	3	3	3.1	3.1	3.3	3		
Industry	4.1 4.3		4.3	4.3	4	4.2 4.3		4.2 4.3	4.3	
Farms & Forests	5.1 5.3 5.4				5.2			5.4		
CO_2 Removal			6.3				6.3	6.3		6.2
Climate Prediction	7.1		7				7.3		7	
Societal Impacts	8.1 8.4	8.4	8.2 8.3		8.2	8.3	8.2	8.1	8.3	
Solar Geoengineering			9.3		9.4		9.3 9.4			9.2
Tools for Individuals	10.1	10.1	10.2	10.3	10.2	10.1			10.2	10.2
Tools for Society		11.1	11.2 11.1	11.3	11.2 11.1	11.1 11.3	11.1	11	11.1	11.1 11.3
Education		12.2			12.1					
Finance		13.2	13				13.2			

Claim Analysis

- Claim verification datasets
  - A claim is potentially verifiable if it is a) **well-formed** and b) **subjectively investigable**
  - CLIMATE-FEVER (Diggelmann et al., 2020)
- Claim verification tools
  - Wang et al. (2021)
  - Bhatia et al. (2021)
- Claim detection
  - Stammbach et al. (2022)

Document Analysis

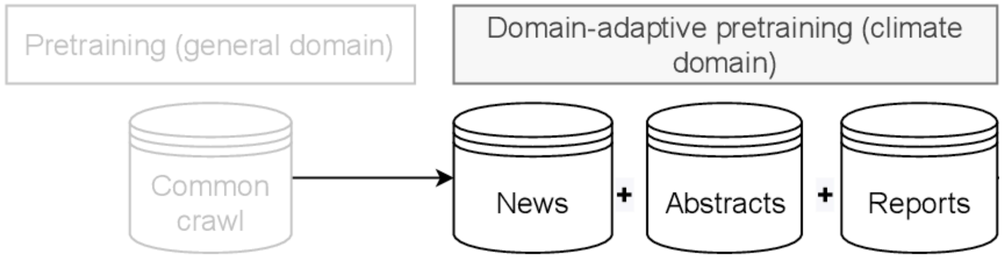
- Topic detection
  - ClimaText (Varini et al., 2020)
- Climate report analysis
  - Bingler et al. (2021)
  - Friederich et al. (2021)
- Generation of evidence maps
  - Callaghan et al. (2021)

Social Media Analysis

- Sentiment, emotion and opinion analysis
  - Jiang et al. (2017)
  - Loureiro & Alló (2020)
  - El Barachi et al. (2021)
- Identification of neutralization and denial
  - Chen et al. (2019)
  - Bhatia et al. (2021)

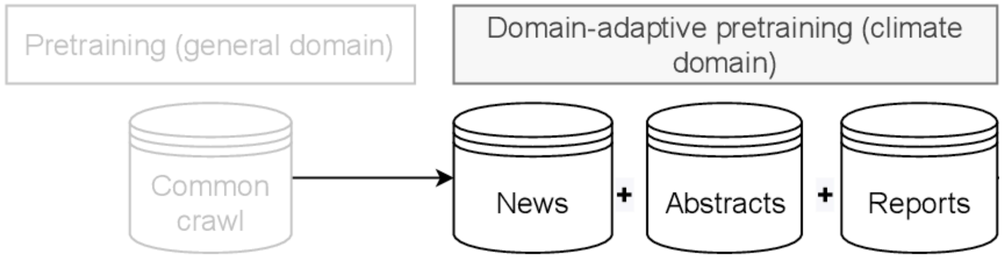
General Purpose Tools

- ClimateBERT (Webersinke et al., 2022)
  - A pre-trained models trained on climate related texts



General Purpose Tools

- ClimateBERT (Webersinke et al., 2022)
  - A pre-trained models trained on climate related texts



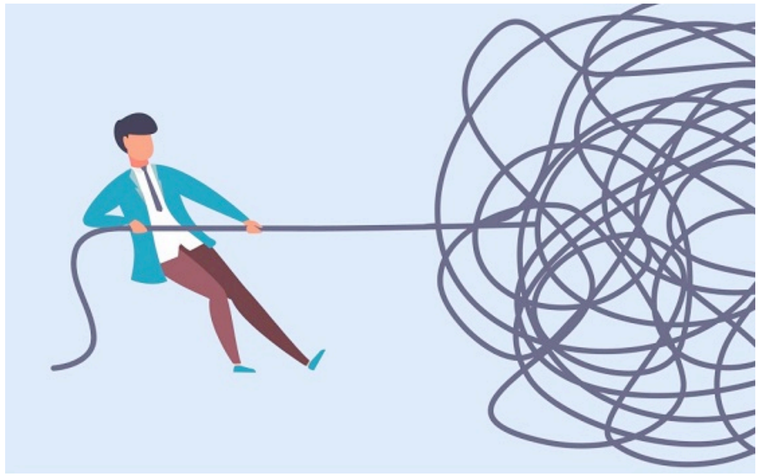
- ClimateQA (Luccioni et al., 2020)
  - A question-answering dataset focusing on climate topics

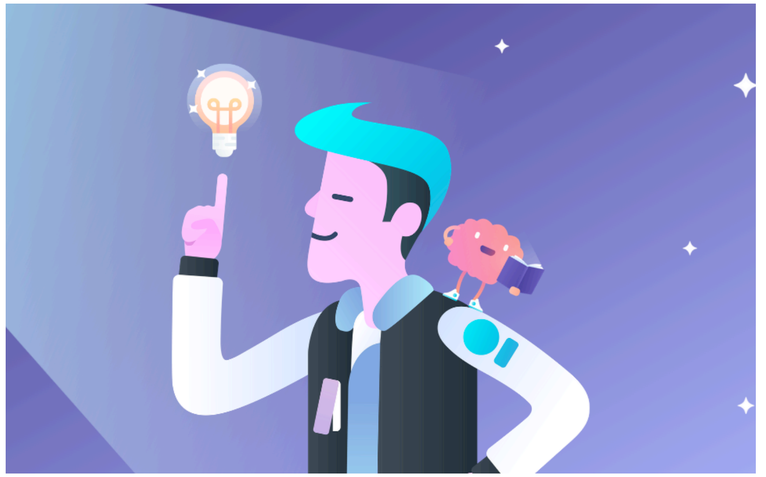
TCFD Question	Answer Passage
Does the organization describe the board's (or board committee's) oversight of climate-related risks and/or opportunities?	*The Company's Audit Committee has the delegated risk management oversight responsibility and receives updates on the risk management processes and key risk factors on a quarterly basis.*
Does the organization describe the climate-related risks or opportunities the organization has identified?	*The availability and price of these commodities are subject to factors such as changes in weather conditions, plantings, and government policies*

How can AI Inform Policy Makers?

- Supporting policy-makers and activists
  - Stede & Patz (2021)
- A knowledge platform for holistic and effective climate action
  - Swarnakar & Modi (2021)

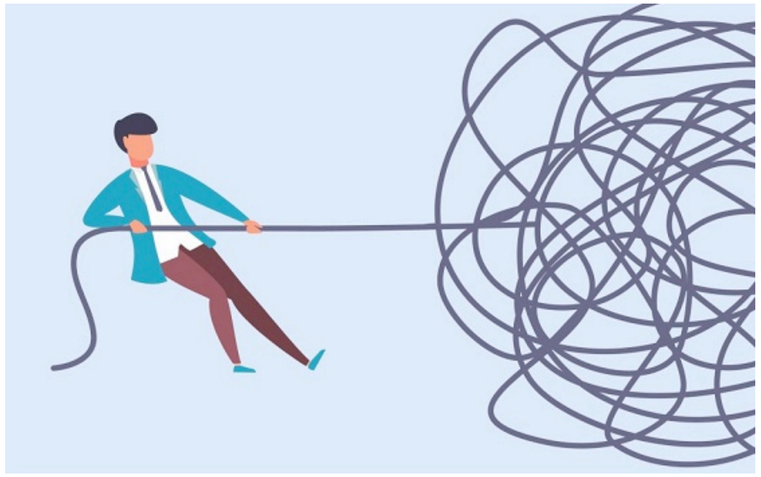
Summary

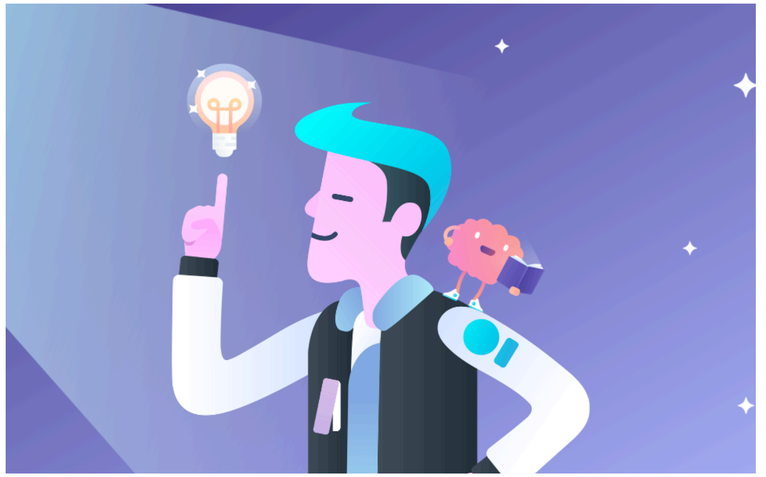




Challenges **Opportunities** **Mitigation**

Summary





Challenges **Opportunities** **Mitigation**